

How Friendly are the Natives?

An Evaluation of Native-speaker Judgements of Foreign-accented British and American English

Published by
LOT
Janskerkhof 13
3512 BL Utrecht
The Netherlands

phone: +31 30 253 6006
fax: +31 30 253 6000
e-mail: lot@let.uu.nl
<http://www.lot.let.uu.nl/>

Cover illustration: by *Marieke van den Doel*

ISBN-10: 90-78328-09-6
ISBN-13: 978-90-78328-09-4

NUR 632

Copyright © 2006: Rias van den Doel. All rights reserved.

How Friendly are the Natives?

An Evaluation of Native-speaker Judgements of Foreign-accented British and American English

De tolerantie van moedertaalsprekers:

een beoordeling van reacties op

Brits en Amerikaans Engels met een buitenlands accent

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht

op gezag van de rector magnificus, prof.dr. W.H. Gispen,

ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen

op donderdag 9 november 2006 des middags te 12.45 uur

door

Willem Zacharias van den Doel

geboren op 11 februari 1965 te Amsterdam, Nederland.

Promotor: Prof. dr. W. Zonneveld.

CONTENTS

Acknowledgements	XI
1 Introduction	1
1.1 Introduction	1
1.2 Background and previous research	4
1.2.1 Native-speaker evaluations of non-native speech	4
1.2.2 Hierarchy of error	6
1.2.3 The effects of different variables on hierarchy of error	10
1.2.4 “English as an International Language” and other approaches	15
1.3 Objectives and methods of the present study	21
1.3.1 General and practical objectives	21
1.3.2 Overview of the methods used	25
1.3.3 A note on statistics	28
2 Design and set-up of the two experiments	31
2.1 The Dutch Experiment: design, subjects and procedure	31
2.1.1 General aims and target groups	31
2.1.2 Sections included in the survey	34
2.1.3 Errors and distractors included in the survey	36
2.1.4 Priorities in data analysis	47
2.2 Dutch respondents’ views on teaching English pronunciation	47
2.2.1 Data analysis and discussion of the results	47
2.2.2 Preliminary conclusions and recommendations	55
2.3 Pronunciation errors in the Dutch Experiment: data analysis	57
2.4 The Native-speaker Experiment: design, subjects and procedure	65
2.4.1 General aims and target groups	65
2.4.2 Sections included in the survey	70
2.4.3 Audio stimuli used in the Native-speaker Experiment	76

2.5	The Native-speaker Experiment: data processing	86
2.5.1	Analysis and categorisation of accent self-identifications	86
2.5.2	Analysis and categorisation of error assessments	100
3	The Native-speaker Experiment: results and analysis	103
3.1	Introduction and overall assessment of leniency and severity	103
3.1.1	Introduction	103
3.1.2	Self-identified leniency	104
3.1.3	Self-identified leniency by sex, age, version and major accent group	105
3.1.4	Self-identified leniency by minor accent group	105
3.1.5	Overall error severity assessment	106
3.1.6	Overall error severity assessment by sex, age, version and major accent group	106
3.1.7	Overall error severity assessment by minor accent group	108
3.2	Overall assessment of individual tokens	109
3.2.1	Overall assessment	109
3.2.2	Hierarchy of error	110
3.2.3	Overall assessment of individual tokens by version of the experiment	115
3.2.4	Overall assessment of individual tokens in the RP version	116
3.2.5	Overall assessment of individual tokens in the GA version	122
3.2.6	Overall assessment of individual tokens: differences between versions	128
3.3	Assessment of individual tokens by age and sex	130
3.3.1	General overview of the results	130
3.3.2	Assessment of individual tokens by sex	131
3.3.3	Assessment of individual tokens by age	133
3.4	Assessment of individual tokens by major accent group	134
3.4.1	Introduction and general overview of the results	134
3.4.2	Pairwise comparisons between GB/RP, US/GA, GB/NRP and US/NGA	138
3.4.3	Respondents' comments on individual tokens	142
3.4.4	Overall effect of detection success on severity	145

3.5	Token-by-token analysis	153
3.5.1	Assessment of BED	153
3.5.2	Assessment of BAT	155
3.5.3	Assessment of VAN	157
3.5.4	Assessment of WINE	159
3.5.5	Assessment of THIN, AUTHOR and BOTH	160
3.5.6	Assessment of THAT, WEATHER and BREATHE	165
3.5.7	Assessment of OFF	169
3.5.8	Assessment of RED	171
3.5.9	Assessment of ICE	173
3.5.10	Assessment of TIE	175
3.5.11	Assessment of DEAD	178
3.5.12	Assessment of FILM	180
3.5.13	Assessment of CAR	183
3.5.14	Assessment of HOT_TEA	186
3.5.15	Assessment of INDIA	188
3.5.16	Assessment of NEW	189
3.5.17	Assessment of PERFECT and IMAGIN	192
3.5.18	Assessment of TO_WALES, THAT_THA and WOULD_ON	195
3.5.19	Assessment of SECONDAR	200
3.5.20	Assessment of TELL	202
3.5.21	Assessment of COLOUR	205
3.5.22	Assessment of STOOD	207
3.5.23	Assessment of INT1, INT2, INT3	210
3.6	Comparison with severity assessment in the Dutch Experiment	217
3.7	General discussion and preliminary conclusions	231
4	Accent Similarity	247
4.1	Token similarity to Dutch English	247
4.2	Similarities to accents in the RP version of the experiment	248
4.2.1	BED	248
4.2.2	BAT	249
4.2.3	VAN	250
4.2.4	WINE	250
4.2.5	THIN, AUTHOR, BOTH, THAT, WEATHER, BREATHE	251
4.2.6	OFF	252
4.2.7	RED	252
4.2.8	ICE	252
4.2.9	TIE	253
4.2.10	DEAD	253

4.2.11	FILM	253
4.2.12	CAR	254
4.2.13	HOT_TEA	256
4.2.14	NEW	256
4.2.15	IMAGIN	257
4.2.16	PERFECT	257
4.2.17	TO_WALES, THAT_THA, WOULD_ON	258
4.2.18	SECONDAR	258
4.2.19	TELL	259
4.2.20	COLOUR	260
4.2.21	STOOD	260
4.2.22	INT1, INT2, INT3	261
4.3	Accent similarity codes for the RP version of the experiment	261
4.4	Similarities to accents in the GA version of the experiment	265
4.4.1	BED	265
4.4.2	BAT	265
4.4.3	VAN	267
4.4.4	WINE	267
4.4.5	THIN, AUTHOR, BOTH, THAT, WEATHER, BREATHE	268
4.4.6	OFF	270
4.4.7	RED	270
4.4.8	ICE	271
4.4.9	TIE	271
4.4.10	DEAD	271
4.4.11	FILM	271
4.4.12	CAR	272
4.4.13	HOT_TEA	273
4.4.14	NEW	273
4.4.15	IMAGIN	274
4.4.16	PERFECT	275
4.4.17	TO_WALES, THAT_THA, WOULD_ON	275
4.4.18	SECONDAR	275
4.4.19	TELL	276
4.4.20	COLOUR	277
4.4.21	STOOD	277
4.4.22	INT1, INT2, INT3	278
4.5	Accent similarity codes for the GA version of the experiment	278
4.6	Accent similarity: analysis and results	280

5	Conclusions	285
5.1	Overview of the analysis of the two experiments	285
5.2	Summary of the main findings	289
5.2.1	Hierarchy of error: general principles	289
5.2.2	Hierarchies of error for the RP and GA versions	291
5.2.3	Detection as a factor in error assessment	296
5.2.4	Error assessment in the different accent groups	298
5.2.5	Accent similarity	301
5.2.6	Comparison with the Dutch Experiment	302
5.3	Limitations of the present study	304
5.3.1	Introduction	304
5.3.2	Presentation of the stimuli	304
5.3.3	Selection of participants	305
5.3.4	Error detection	306
5.3.5	The Dutch Experiment	307
6	Recommendations	309
6.1	Recommendations for future research	309
6.2	Implications for teaching English in the Netherlands and elsewhere	311
6.2.1	Introduction	311
6.2.2	Beginners' or intermediate level	313
6.2.3	Advanced level	316
	References	319
	Summary in Dutch	335
	Curriculum Vitae	341

ACKNOWLEDGEMENTS

Three people have been indispensable for bringing this dissertation into existence and seeing it through to a long-awaited conclusion: firstly, Peter Coopmans, for encouraging me to begin; secondly, Wim Zonneveld, for agreeing to act as my supervisor; and thirdly, Bev Collins, for offering to serve as my Welsh phonetics “gwrw”. Without Peter’s insistence, Wim’s constant involvement, unfailing support and stimulating advice, and Bev’s generous help, this dissertation could not have been written. I am most grateful to Bev for his unstinting commitment to removing all my non-native errors, as well as all my jokes, from the text.

I also owe a debt of gratitude to the other members of the reading committee: Hugo Quené, Vincent van Heuven and René Kager, both for their comments and their help with earlier incarnations of the manuscript. If it had not been for Hugo’s unflagging support with the statistical side of things and his kind words of encouragement, I would probably still be entering meaningless data into Excel files. I should like to thank Vincent for his time and trouble in fine-tuning the recordings for the Native-speaker Experiment, and his significant involvement in recording, transcribing and analysing the intonation sentences. In addition, I wish to acknowledge René’s help in bringing me up to speed on some of the secondary literature in question, of which I had been blissfully ignorant until that time.

I could not have designed the two online experiments without the truly essential help of Theo Veenker, who built the WWStim software, helped me create the relevant hypertext, and assisted me with the recordings and the data processing. I am grateful to Guus de Krom for offering invaluable advice on experimental design and statistics, and for guiding me through the early stages of my research. I should also like to acknowledge the assistance freely given by Paul van Buren and Robert Lankamp, who lent their voices to the Native-speaker Experiment and gracefully complied with the antics I put them through.

My thanks go out to all those many people who agreed to do one of my experiments. Among these are all the secondary school teachers, students of English, college and university lecturers in the Netherlands who took part in the Dutch Experiment, including my own students, colleagues and friends. Furthermore, I would like to thank all the English speakers from around the globe who participated in the Native-speaker Experiment – including those who commented on the first pilot version. Many respondents forwarded my call for participants to their families, friends and colleagues – providing the “snowball effect” described in 2.4.1. In addition, I would like to thank those participants who sent me lengthy emails concerning the experiment.

Over the last few years, many other people have given me books and articles to read, or advice and help in various other ways. These include Theo Bongaerts, Rachel Collins, Jeroen van Dordrecht, Miriam Ebsworth, Esther Janse, Janet Grijzenhout, Monique van der Haagen, Keetje van den Heuvel, Bregje Holleman, Ton Koet, Klaske van Leijden, Inger Mees, Frank van Meurs, Debi Molnar, Martina Noteboom, Wim Peeters, William Philip, Greetje Reeuwijk, Alyson Ross, Bert Schouten, Maria Sherwood-Smith, Joost Uding, Sharon Unsworth, Hans van de Velde and Wim van der Wurff. I owe them a debt of gratitude. In particular, I would like to acknowledge the invaluable help of Swedish librarian and friend Katarina Standár, who managed to track down a great many articles which would otherwise have been unobtainable.

In addition to those who commented on earlier versions of the manuscript, or parts of it, I would also like to thank two people who patiently and painstakingly proofread the entire text: Maxim Brouwer and Colin Ewen. I cannot begin to thank them enough. I will never again assume that all dashes are the same, or that *Word* can be relied upon to arrange lists in alphabetical order. Needless to say, any remaining errors are the result of my own stubbornness or inconsistency [sic!].

I am grateful to the Faculty of Arts at Utrecht University, and the Utrecht Institute for Linguistics, for funding my research trips to Britain, and for paying for the book tokens presented to selected participants in my experiment. I would also like to thank my colleagues and students at Utrecht and Leiden University for their forbearance, support and company when I was engaged in writing this dissertation – including helping me move office or taking on some of my workload. In this respect, I would like to express my indebtedness to Simon Cook, Onno Kusters and Roselinde Supheert, and other colleagues mentioned above.

Finally, I wish to thank my family and friends for their love and constant support – my parents above all. I am particularly grateful to my sister Marieke for drawing the cartoon on the cover of this book. My greatest debt of gratitude, however, is to my husband Maxim. He knows why. To him I dedicate this book.

CHAPTER 1

INTRODUCTION

1.1 Introduction

This dissertation investigates the notion of foreign accent. More specifically, it investigates the evaluation of aspects of foreign-accented English by different groups of native speakers, and does so by means of empirical-experimental methods.

Foreign accents are commonly subject to stereotyping by native speakers.¹ This is widely exploited in film and television as, for instance, is demonstrated by Lippi-Green's (1997) study of how animated Disney films employ a range of accents to "perpetuate stereotypes on the basis of language" (Lippi-Green 1997: 101). What is perhaps less well-known is that learners of a foreign language also have certain prejudices about the way their speech is evaluated by native speakers of that language. For instance, a majority of Dutch students of English appear to believe that English native speakers from Britain and Ireland are the most severe judges of their Dutch-accented pronunciation of English; only a small minority consider North Americans to be less lenient. This is apparent from a brief web survey conducted by the author of this thesis in the Netherlands in June 2005, in which 615 Dutch participants were asked which groups of English speakers they believed were the strictest judges of Dutch pronunciation errors. There were four response categories: (1) "inhabitants of the British Isles" (2) "Americans and Canadians" (3) "other groups of native-speaker judges" (4) "don't know". The survey, conducted in Dutch, was carried out with students enrolled for a degree course in English at the University of Utrecht, together with their friends, relatives and online contacts.

As Table 1.1 shows, while 56% of respondents selected the "British and Irish" group, only 7% opted for the North Americans. While 9% opted for "Other" (which, as the comments revealed, included Dutch teachers of English), 28% stated that they did not know. The differences between the four options were highly significant ($\chi^2 = 389.2$, $df = 3$, $p < .0001$). Since it was assumed that students who had taken a course in English phonetics would be more aware of pronunciation issues, which could affect their response, participants were asked

¹ In this study, the term "accent" will be used to denote "[a] particular way of pronouncing a language, seen as typical of an individual, a geographical region, or a social group. Every speaker of a language necessarily speaks it with some accent or other" (Trask 1996: 4). As Richards *et al.* (1985: 1) point out, this could refer to "the region or country", "the social class [the speakers] belong to", and "whether or not the speaker is a native speaker of the language".

whether or not they had taken such a course at university or college level. This, however, did not affect their answers significantly ($\chi^2 = .642$, $df = 3$, n.s.). These results clearly suggest a large degree of consensus amongst Dutch respondents, regardless of whether or not they had studied phonetics, that British and Irish people are more stringent judges of Dutch pronunciation errors in English than are any other groups.

Table 1.1. Frequencies and percentages of Dutch respondents who answered to question “Which of the following groups are the strictest judges of Dutch pronunciation errors in English?” by response category. Respondents have been divided into those who studied English phonetics at a Dutch university or college and those who had not.

	British and Irish	Americans and Canadians	Other	Don't know	Total
Phonetics students	153 (57%)	20 (7%)	24 (9%)	70 (26%)	267 (100%)
Other	194 (56%)	24 (7%)	29 (8%)	101 (29%)	348 (100%)
All respondents	347 (56%)	44 (7%)	53 (9%)	171 (28%)	615 (100%)

As will be demonstrated by the experimental research discussed in the present study, British and Irish speakers of English are certainly not the strictest judges of Dutch pronunciation errors in English. As is shown in Chapter 3, it is the speakers of US and Canadian English who evaluate most severely the errors they detect in the stimuli presented to them. In other words, the perceptions which many Dutch learners have about the relative leniency of different groups of native-speaker judges are inaccurate, and are likely to be based on stereotypical notions of these groups and their cultures. In this case, the tendency to view English speakers from the “British Isles” as stricter judges may be affected by the fact that the implicit norm for English teaching in most Dutch secondary schools is a variety of British English (cf. Van der Haagen 1998: 2).² Such perceptions are likely to affect learners’ attitudes and motivations. It may cause some Dutch learners to believe, for instance, that their foreign accents are not subject to any form of criticism by North Americans. The latter is a very serious misconception, as is evident from the relatively strict evaluations of Dutch pronunciation errors by American and Canadians (discussed in Chapter 3).

² Some people in Ireland consider the term “British Isles” offensive, as it could erroneously imply that Ireland is part of Britain.

There have been numerous studies of native-speaker reactions to foreign accents, an overview of which is presented in 1.2.1. Very little research, however, has been carried out on the way that non-native speech is evaluated by native speakers who speak different varieties of the same language, as in the case of British, Irish and North American speakers of English. This is particularly true of experimental research which attempts to establish the relative importance of various pronunciation problems of foreign learners in what has been termed a “hierarchy of error”. It is also an important objective of the core experiment discussed in this dissertation to devise such an error hierarchy for Dutch learners of English. What is new about this experiment, however, is that it is the first to compare and contrast the evaluations of discrete features of non-native speech by different groups of native speakers on a large-scale, structural basis. It is also the first to do so using appropriate statistical methodology (for a description of the latter, see 1.3.3).

The main goals of the core experiment of this dissertation are, firstly, to investigate how various groups of native speakers prioritise certain features of Dutch-accented English, and, secondly, to capture these prioritisations in a number of different hierarchies of error. To this end, respondents were drawn from Britain, Ireland, the United States, Canada, Australia, New Zealand and South Africa. It was assumed that, if it could be established that native speakers of English with different linguistic and cultural backgrounds in fact evaluate and rank Dutch pronunciation errors in English differently, this would have important implications in a number of areas for teaching and research. For instance, it would suggest that a stronger emphasis on a sociolinguistic and variationist framework would be required not only for research into native-speaker attitudes to foreign accents, but also into second language acquisition (as advocated by Bayley 2000 and Bayley & Regan 2004), and, in addition, with regard to EFL pronunciation training as a discipline.

If native-speaker judgements of foreign accents are shown to be affected by accent variation in the judges themselves, this would demonstrate that native-speaker norms are neither monolithic nor immutable, and that the “overemphasis in S[econd] L[anguage] A[cquisition] on the standard language” (Bayley 2000: 289) is in serious need of revision. As Bayley (2004: 289) puts it, “acquisition needs to be judged not in terms of the standard language, but in terms of the varieties with which learners are in most frequent contact”. Additionally, given the abundance of varieties of English that exist, the unprecedented exposure to these models currently enjoyed by learners through the different media, and the widening scope for learner confusion that this situation could engender, awareness of linguistic variation is a didactic concern that is beginning to be increasingly urgent in English teaching at virtually every level. Furthermore, from a practical point of view, the approach adopted in the present study would make it possible to draw learners’ attention to those foreign pronunciation features detected most readily, and assessed most negatively, by speakers of those accents they may be imitating, as well as making them aware of the different priorities assumed, in this respect, by other groups of native speakers.

After all, a balanced proficiency curriculum should not aim to prepare learners for interactions with one type of native speaker only.

The objectives and methods of the present study will be discussed in more detail in 1.3. However, it may be noted here that a third objective of this study was to discover to what extent factors such as native speakers' sex, age, linguistic or educational background affect their attitudes to particular features of non-native speech. For example, some interlocutors may be less inclined to notice or reject non-native realisations that are perceived as similar or identical to what may be heard in their own speech community (cf. Johansson 1978: 95). One could argue that a native speaker who pronounces *that* with an initial stop is unlikely to object to a similar realisation in a foreign learner. It is a core objective of this dissertation to test the assumption that "accent similarity" will cause the judges involved to be more lenient.

Finally, a significant consideration in this study is that the non-native speech features to be evaluated by native speakers should be based on realistic pronunciation problems which are attested in the actual training of foreign learners. This should make it possible for any conclusions to be drawn from this research to be directly applicable to teaching practice – which is the fourth, more practical, objective of this study. In view of this, it was decided to select a number of representative Dutch pronunciation errors in English from a well-known pronunciation manual (Collins & Mees 2003b) and ask a number of individuals involved in English language acquisition in the Netherlands to pick out the most significant of these. This was then submitted to native-speaker judges, who were subsequently asked to rank these errors according to their severity. It was hoped that this procedure would also allow a comparison between native and non-native evaluations of foreign-accented speech.

1.2 Background and previous research

1.2.1 Native-speaker evaluations of non-native speech

According to some scholars, pronunciation teaching may even be considered immoral. As Porter & Garvin (1989: 8) have argued: "To seek to change someone's pronunciation – whether of the L1 or of an L2 – is to tamper with their self-image, and is thus unethical – morally wrong". It may seem surprising that such perceptions exist about pronunciation training for non-native learners. There is a large body of evidence which shows that non-native accents are, generally speaking, subject to negative evaluations by native speakers (for an overview, see Ryan 1983, Eisenstein 1983, Munro & Derwing 1995, Leather 1999, Major *et al.* 2005, Scheuer 2005). As a result of these assessments, "nonnative speakers may be personally downgraded because of their foreign accent" (Leather 1999: 35) and be accorded "a lack of competence in many spheres" (Ryan 1983: 155). Arguably, the self-image of non-native speakers will

be adversely affected much more strongly by such negative evaluations than by any remedial pronunciation training.

The effect of such negative appraisals may be diminished or enhanced by a number of factors, which include the “degree of accentedness” (see Ryan & Carranza 1976, Sebastian *et al.* 1978) and the interaction of this element with speakers’ speech styles and social class background (Ryan & Sebastian 1980). Another consideration is the actual status that some foreign accents have for certain groups of native (and non-native) speakers. In the US, for instance, a Spanish accent in English is, on the whole, more prone to stigmatisation than a German one (Ryan 1983: 154), and may even be assessed negatively by second-generation immigrants from Mexico (Ryan *et al.* 1975). Delamare (1996) found that American listeners viewed speakers with certain foreign accents (such as Arabic and Farsi) more favourably if the individuals made grammatical errors than if they did not, whereas speakers with other accents (such as French and Malay) were actually downgraded if they did produce such errors. This implies that the social context in which native speakers encounter a foreign accent plays an important part in their evaluation of the accent concerned. In particular, when non-native speakers assume “more demanding social roles”, which presuppose a large degree of “public accountability”, the extent of their foreign accent is likely to be scrutinised more closely by natives (Bresnahan *et al.* 2002: 173, based on Cote & Clement 1994). What native speakers will accommodate in a friend they may object to in a professional exchange (Bresnahan *et al.* 2002: 171). If, in such public settings, native speakers of a particular language are dominant, non-native speakers with strong accents may find themselves marginalised and relocated to the periphery (Scheuer 2005: 125–126).

It would be pointless to suggest that, in view of these negative assessments by native speakers, or as a result of some more positive objective, all non-native speakers of a language should seek to eliminate their foreign accents. To begin with, this would be an impossible aim to achieve for the overwhelming majority of adult learners. It is true that there are a few documented cases (e.g. Bongaerts 1998, see also Leather 1999: 10) where post-pubescent learners have managed to achieve completely native-like accents, despite the fact that this took place *after* the “critical period” in which it is often assumed that learners are capable of learning to mimic perfectly the pronunciation of a second language (cf. Lenneberg 1967, Scovel 1988). However, as Bongaerts (1999b: 155) pointed out, “the success of the exceptional adult learners” in question was “at least partly due to the combination of three factors: high motivation, continued access to massive L2 input, and intensive training in the perception and production of L2 speech sounds”. In addition, the native language of these learners (Dutch) was typologically closely related to the relevant target language (English). As Bongaerts *et al.* (2000: 307) proposed, “in the domain of pronunciation ... typological proximity may be one of the determining factors of ultimate nativelike performance”.

Clearly, such conditions cannot be met by the vast majority of learners worldwide. Even in those cases where students are highly motivated, they will

not always have access either to “massive L2 input” or to intensive speech training. (The present author is unaware of any such training being made available, on a large scale, to adult immigrants anywhere in the world.) Moreover, if typological proximity is a crucial requirement for the late acquisition of a native-like accent, this will present an insurmountable barrier, for instance, to most of the world’s adult learners of English (except, presumably, to speakers of other closely related languages).

There are many other reasons why non-natives may retain their foreign accents, and it would go beyond the scope of this study to discuss these in full. However, it may be noted that those lacking an integrative motivation (cf. Ellis 1994: 509–513) are unlikely to aspire to a native-speaker model, and their retention of a non-native accent may, for instance, “occur as an expression of ethnic identity, as an emotional statement of defiance, and as a means of facilitating social categorization” (Ryan *et al.* 1980: 1, based on Giles & Powesland 1975; see also Taylor & Giles 1979). Ryan *et al.* (1980: 1) define the last notion as the “positive influence of designation as a foreigner (or minority individual) on social evaluation within a particular context”. For instance, as Ryan (1983: 157) has argued, “a certain amount of nonstandardness (e.g. a language learner’s accent) can sometimes attenuate the impact of another aspect of nonstandardness (e.g. grammatical or sociolinguistic errors)”. It may be noted, however, that Ryan and her associates also discovered that “the social utility of protecting oneself from the consequences of an inadvertent *faux pas* by retaining a Spanish accent in English would apparently be gained only at the considerable expense of being viewed as less successful, intelligent and wealthy” by Anglo-American listeners (Ryan *et al.* 1980: 6).

It is questionable whether such strategies are deliberately undertaken by members of immigrant communities, who are more likely to retain non-native accents, at least in part, as badges of ethnic identity and as expressions of solidarity with others in the same “in-group” (Ellis 1994: 211). In addition, “negative attitudes towards the target-language culture” (Ellis 1994: 208) and “fear of assimilation by that group” (Ryan 1983: 154) may also play a part. Nevertheless, such factors are much less likely to affect those approaching the target language in an educational setting in which the learner’s native language is associated with the dominant majority (cf. Ellis 1994: 209), as is presumably typically the case with most learners of English on the European continent, which is the target group of this study.

1.2.2 Hierarchy of error

In short, there are a great many reasons why most non-native speakers do not, or cannot, eliminate their foreign accents. The British phonetician Abercrombie (1956) asked the question whether it was “really necessary for most language learners to acquire a perfect pronunciation”, going on to say:

Intending secret agents and intending teachers have to, of course, but most other language learners need no more than a comfortably intelligent pronunciation (and by

“comfortably” intelligible, I mean a pronunciation which can be understood with little or no conscious effort on the part of the listener). I believe that pronunciation teaching should have, not a goal which must of necessity be normally an unrealized ideal, but a *limited* purpose which will be completely fulfilled: the attainment of intelligibility. The learner, instead of being taken through each English vowel and consonant, and later, if there is time, through the complexities of intonation and rhythm, would have presented to him certain carefully chosen features on which to concentrate, the rest of his pronunciation being left to no more than a general supervision (Abercrombie 1956: 93).

What Abercrombie is here recommending in effect is that learners still model their L2 accents on native speakers of the target language in question, but only concentrate on a limited number of pronunciation problems, which are to be carefully selected and prioritised on the basis of their consequences for intelligibility. If it can be established which learner errors are most likely to cause a breakdown of intelligibility, these may be presented to learners in the form of a hierarchy of error, i.e. an overview of those pronunciation problems which merit their attention most – a central notion of this thesis.

Following seminal work by Johansson (1973, 1975) on the notion of hierarchy of error, and his experimental research into the reactions of speakers of British English to Swedish-accented English (Johansson 1975, 1978), various attempts have been made to establish such hierarchies for different groups of learners. For instance, experimental research based on native-speaker judgements was done by Dretzke (1985) to establish a hierarchy of pronunciation errors for the benefit of German learners of English, by Norell (1991) for Swedish learners of English, by Schairer (1992) for English-speaking learners of Spanish, and by Koster & Koet (1993) for Dutch learners of English.

Other similar hierarchies of error have been formulated partly on the basis of experimental research, but mainly as a result of impressionistic observational procedures. These error hierarchies include those put forward in Collins & Mees for Dutch learners of, firstly, British English (2003b: 290–293, originally in Collins & Mees 1981: 196–197, also in modified form in Collins *et al.* 1987) and later for American English (Collins & Mees 1993).³ This approach also seems to be the basis of the list of pronunciation priorities proposed in Gussenhoven & Broeders (1997: 16–17). In the same vein as Collins & Mees (1993, 2003b), the present study aims at establishing error hierarchies for the Dutch pronunciation of both British and American standard varieties of English,

³ Collins (1979a, b) conducted a pilot experiment investigating the hypothesis that English native speakers and Dutch teachers of English would arrive at different error hierarchies for pronunciation. The preliminary results, which confirmed this hypothesis, were presented at the 1978 10th International IATEFL conference in London and also at the 1979 2nd International Teaching of Spoken English Conference at Leeds University. The project was subsequently shelved, but the native-speaker reactions formed a basis for the hierarchy of error to be found in Collins & Mees (1981, 2003b) and Collins *et al.* (1987).

although employing only experimental research. In this respect, the present study is akin to Dretzke's (1985) attempt to devise a hierarchy of error for German speakers of English, and also to a similar study undertaken by Koster & Koet (1993) on hierarchy of error for Dutch learners of English. Perhaps the chief way in which this dissertation differs from Collins & Mees (2003b, 1993), Dretzke (1985), Koster & Koet (1993) and all other such studies is the employment of an Internet-based enquiry to determine native-speaker reactions.

Most other attempts to establish hierarchies of error (i.e. in addition to those above) have not focused exclusively on pronunciation, but have sought, for instance, to establish the relative importance of pronunciation errors as against other types of error (for an overview, see Johansson 1978: 9–15, Ludwig 1982, Eisenstein 1983: 163–168, Fayer & Krasinski 1987: 314–315, Munro & Derwing 1995: 75–76, Rifkin 1995: 477–478, and the references contained therein). Given the differences between the experimental design, the different L1 and L2 languages concerned and the types of error under examination, it is hardly surprising that these hierarchies cannot be reliably compared. However, it is interesting to note that one such study (Albrechtsen *et al.* 1980) of native-speaker judgements of Danish errors in English concluded that a hierarchy of error cannot actually be established, as the context in which the errors occurred played an important part in their effect (see 3.7 and 5.1 for discussion).

It should be pointed out that in most studies of error hierarchy, intelligibility is not seen as the only factor in determining error gravity, i.e. the significance of a learner's error as perceived by native speakers. An important principle of evaluation, as noted by Johansson (1978: 4), is that even if “the erroneous utterance is fully comprehensible, it could nevertheless have serious consequences from the point of communication, e.g. make the receiver tired or irritated or draw away his attention from the contents of the message”. The importance of “distraction/irritation on the part of the native-speaker listener” is also emphasised by Collins (1979b: 27). Arguably, such irritation may partly account for the generally negative evaluations of foreign-accented speech.

In their study of Mandarin-accented English as evaluated by speakers of Canadian English, Munro & Derwing (1995) also found that strongly accented speech cannot be equated with a lack of intelligibility, in spite of the correlation they discovered between these factors. (Perhaps somewhat paradoxically, they concluded from this that pronunciation training should concentrate on reducing unintelligibility.) Scheuer (2005: 116) draws the same conclusion that “foreign accent and unintelligibility are not synonymous” and refers to research by Markham (1997: 101) into native-speaker reactions to L2 Swedish which showed that “the more negatively judged errors are ones which *do not* cause lexical confusion ... – they are simply non-native pronunciations – , whereas the more acceptable errors *can* cause lexical confusion” (italics Scheuer's).

If foreign accents are downgraded in spite of their intelligibility, this implies that non-native speakers should also avoid pronunciation features which may cause their native-speaker interlocutors to be distracted or irritated. For instance, Cunningham-Andersson (1997) investigated a number of virtually

identical pronunciation errors made by different speakers of L2 Swedish with various linguistic backgrounds and found that some of these errors were more stigmatised than others. She concludes that “[i]t would clearly be worthwhile for immigrants to learn to avoid the non-native pronunciations that are stigmatized” (Cunningham-Andersson 1997: 142).

Interestingly, a study by Piazza (1980: 424–426) into the reactions of French secondary school pupils to grammatical mistakes made, in both speech and writing, by American learners of French revealed that “[i]rritation was judged more severely than lack of comprehensibility”, especially in spoken language (cf. Ludwig 1982: 275). At the same time, Piazza (1980: 424) also found that the greater the loss of intelligibility resulting from a particular type of error, the more irritating it was considered to be. As Fayer & Krasinski (1987: 315) have argued, this “negative correlation between the degree of irritation and the degree of unintelligibility” suggests that irritation can actually be “the possible result of unintelligibility”. In other words, the degree of native speakers’ irritation with strongly accented speech may be partly dependent on their ability or inability to understand the message.

Clearly, it is difficult to separate out the effects of unintelligibility and irritation. This is why it should not be assumed that learners should concentrate only on those errors which are likely to cause intelligibility breakdown. As Johansson (1978: 6) points out, “communicative efficiency does not mean comprehensibility in the strict sense. Speech can be severely distorted and yet be intelligible, as is shown by numerous experiments ...”. Johansson (1978: 6) goes on to say that, “[t]o be communicatively effective, the message must get across swiftly and unambiguously and without undue demands upon the receiver”. Despite the position taken on this by Munro & Derwing (1995: 93), “mere” intelligibility does not suffice to ensure efficient communication.

In fact, native speakers may be perfectly capable of processing certain L2 errors whilst simultaneously considering these unacceptable, as has been shown, for instance, in two studies by Guntermann (1978) and Chastain (1980) of native-speaker evaluations of L2 grammatical errors by English-speaking learners of Spanish. According to Ludwig (1982: 278), both studies suggest that “if the goal of the L2 learner is to establish social and personal relations with N[ative] S[peaker]s, certain errors may be more stigmatizing than others” (cf. Guntermann 1978: 252). Chastain (1980: 214) speculates that this may be particularly true of very basic errors, in which case the “commonality and the simplicity of the pattern make it very difficult for native speakers to sympathize” with “error prone non-natives”.

Chastain’s comment appears to be similar to Johansson’s repeated suggestion (1975: 22–29, 1978: 6–7) that “generality” should also be a concern in error evaluation. As Johansson (1978: 6–7) states, “[a]n error involving a general rule reveals a weakness that may affect an infinite number of cases and may therefore have more serious consequences for communication than errors involving individual items (words or grammatical exceptions)”. Another such general principle invoked by Johansson (1978: 6–7) in this context is the

“frequency” of error, as “an error involving frequent words or constructions may affect a larger number of cases in actual communication” (cf. Johansson 1975: 22–29). As Ellis (1994: 66) points out, irritation and frequency may also be interrelated.

The gravity of an error is thus not simply based on intelligibility, but considerations such as irritation, acceptability, generality and frequency also appear to play a role in the severity of an error’s assessment by native speakers. In view of the likelihood of there being interaction between such factors, one wonders to what extent it is realistic to attempt to distinguish very precisely between them, as appears to be suggested by Johansson’s (1975: 26–29, 1978: 4–7) somewhat elaborate disquisition on devising a hierarchy of error. This would suggest that, while the importance of such aspects may be acknowledged in the selection of the errors, and in the discussion of their significance as indicated by respondents, it may not be useful to ask the latter to rate these elements independently of each other. Johansson (1975: 31) has also pointed out the difficulties of measuring “the receiver’s irritation directly”, and has suggested that “the reactions observed could be taken as overt indications of the disturbing effect of the errors”. With regard to the other points mentioned above, Johansson (1978: 7) argues that the “relative importance of these principles of evaluation may not be the same for all kinds of learners but varies depending upon such factors as the goal of the studies and the stage of learning (cf. Johansson [1975]: 22ff)”. It would be difficult to require native-speaker respondents to consider such aspects separately in their evaluations of non-native speech.

1.2.3 The effects of different variables on hierarchy of error

Attempts to establish a hierarchy of error may be further complicated by the possible effect of other variables. As Johansson (1975: 31) puts it:

Establishing the communicative effect of learners’ errors is no easy task, since it may be assumed to vary depending on such factors as the type of speech situation, the receiver’s age and educational level, general psychological characteristics of the receiver and the degree of their association with foreigners, etc. All of these factors should be kept in mind in an exhaustive study.

Similarly, Gass & Varonis (1984: 81) found that a native speaker’s familiarity with “the topic of discourse”, with “nonnative speech in general”, with “a particular nonnative accent” and with “a particular nonnative speaker” all facilitate “the native speaker’s comprehension of nonnative speech”. This implies that communicative efficiency is reduced if familiarity with any of these four elements is absent. This effect is likely to be particularly strong if some of these features are combined – something to be borne in mind especially by speakers of lesser-known languages addressing a general audience of foreign monolinguals on a specialist topic for the first time (such as a Dutch expert in some field speaking in English on American television).

Whereas a significant number of studies have researched the effects of such indexical, sociolinguistic and psychosocial factors on native-speaker evaluations of non-native speech, very few have investigated their significance in establishing a hierarchy of error for pronunciation. In studies of hierarchy of error not exclusively devoted to speech, often only a general reference is made to such aspects. For instance, in his research into German evaluations of errors made by English speakers of German, Politzer (1978: 256) states: "We do not know to what extent sociological, educational, and above all, also German dialect variation among the subjects may influence the evaluation of specific items or whole categories of errors". Similarly, Piazza (1980: 426) warns the readers that her conclusions about French evaluations of grammatical errors made by Americans are only based on a population of "seventeen- and eighteen-year-old Parisian students" who acted as judges, and goes on to suggest that "[f]urther studies might investigate reactions from different segments of the French population". However, Rifkin (1995) provides a more specific comparison of the evaluations of a number of errors made by American learners of Russian by non-native teachers of Russian and native speakers of Russian respectively, and finds that there is broad agreement between these groups (except in the case of accurate grammatical gender, the importance of which the non-native teachers appear to underestimate). Interestingly, Rifkin (1995: 488) attributes this to attempts on the part of the instructors to "assume the perspective of a native noninstructor in assessing the communication skills of their students", as a result of which "they are likely to respond relatively accurately to the various successes and failures they encounter in their students' spoken Russian".

Such comparisons of the different evaluations of non-native language output by native and non-native judges, and by instructors and non-instructors, are not uncommon in the relevant literature. An overview of these is presented in Ellis (1994: 63–67; see also 3.6). Unlike the investigation by Rifkin (1995), these studies show a general tendency for non-native judges and instructors to evaluate foreign learners' errors considerably more severely than do native speakers and non-instructors. For instance, Koster & Koet (1993: 69) compared the different ways in which Dutch non-native teachers of English and English native speakers evaluated Dutch-accented English and found that the former were stricter than the latter, possibly as a result of "undue fastidiousness" (Koster & Koet 1993: 69). A similar severity in non-native teachers of English composition was also attested in Hughes & Lascaratou (1982) and Sheorey (1986). Likewise, Galloway (1980) and Schairer (1992) found that native speakers who did not teach Spanish were more lenient judges of English-accented Spanish than those who did. Fayer & Krasinski (1987: 321) also found that Spanish-speaking judges of Puerto Rican-accented English were stricter than native speakers of English, and suggested that "nonnatives, no matter what their proficiency level, are embarrassed by their compatriots' struggles in the nonnative language".

There are also studies which do not provide any evidence for the claim that non-natives, and teachers in particular, are more severe judges than native speakers. For instance, two experiments by Johansson (1978: 128) comparing the evaluations of Swedish English by native speakers of English and Swedish respectively did “not support the alleged greater tolerance of overt errors among native speakers”. He did, however, find that the former attached more significance to prosody rather than segmental errors (Johansson 1978: 9–15, 123), a result which was not replicated by Koster & Koet (1993, see 3.6 for details). A tendency on the part of non-native speakers to prioritise different *types* of errors, for example “global” errors that “affect overall sentence organization” has also been attested in other studies (cf. Ellis 1994: 66, Dulay *et al.* 1982: 191). It may also be noted that Bongaerts (1999a: 9) actually found that non-native speakers were less reliable judges of pronunciation than native speakers when it came to identifying learners as non-native. This was irrespective of whether they had experience of judging or teaching pronunciation. In fact, in an overview of studies of the different acceptability judgements of respondents classed as linguistically “naive” as opposed to “sophisticated”, Johansson (1978: 22) found that “linguistic sophistication may be an obstacle rather than an advantage in judgements of acceptability”.

Much less research has been done into the effects of sex and age of the judges in their prioritisation of non-native pronunciation errors. A possible reason why so little work has been done on the influence of age on error tolerance is that such studies tend to rely on the participation of pupils and students.⁴ Be that as it may, neither sex nor age appear to be significant factors in pronunciation hierarchies such as those established by Johansson (1978), Dretzke (1985) or Koster & Koet (1993). Nevertheless, in her study of native-speaker reactions to Spanish pronunciation errors produced by English learners, Schairer (1992: 311) observes that female judges were “found to be more strict than their male counterparts in the evaluation of comprehensibility, particularly at the lower performance levels”, whereas the male judges were “marginally more strict at the upper levels”. Additionally, in her review of studies of native-speaker reactions to non-native errors, Ludwig (1982: 280) concludes that “[y]ounger informants and those who have undergone less rigorous academic programs tend to be more accepting than their opposites of errors of all types”, whereas Eisenstein (1983: 166) cites one example of younger students who were stricter than adults in judging gender errors in French, possibly “reflecting a normative attitude associated with the prescriptive orientation of the school environment”. It has been suggested that younger judges are less experienced with “language variations” (Ryan 1983: 154) and therefore possibly more intolerant, and that older judges may be less strict as a result of their greater exposure

⁴ A notable exception is Albrechtsen *et al.* (1980), which employed 120 adult native speakers as well as 180 pupils aged 16 to 17. The statistically significant differences attested between these groups do not appear to be relevant to a hierarchy of error, and are therefore outside the scope of this dissertation.

to language variation (cf. Major *et al.* 2005: 45). In other words, the influence of education and experience on younger and older judges seems to be a moot point.

Virtually no research has been carried out on the way that judges with the *same* native language but with *different* social or regional accents evaluate foreign learners' pronunciation errors. There are, of course, numerous studies that have investigated the attitudes of such disparate groups of judges to other native speakers of the same language. A classic example is Labov's (1966) study of the New York City accent, which included an experiment in which native speakers of English with different social and ethnic backgrounds were asked to evaluate the career opportunities of a number of fellow New Yorkers from Manhattan's Lower East Side on the basis of their speech (Labov 1966: 405–454). There have been far fewer investigations into the reactions of such differently accented native speakers to *non-native* speech. For instance, Johansson (1978: 9) describes research by Bansal (1965/66, 1969) which showed that British and American respondents understood the English of Indian learners "equally well" (and significantly better than did German or Nigerian listeners). A number of other similar studies (e.g. Albrechtsen *et al.* 1980, Bresnahan *et al.* 2002) have also compared, implicitly or explicitly, the attitudes of differently accented native speakers to foreign accents.

Interestingly, in the vast majority of such investigations, no distinction is made between any differences that may exist between the accents of the judges themselves. Most studies of attitudes to non-native speech do not provide a basis for comparison between native speakers from different countries, as respondents, in their capacity as students or teachers, are normally drawn from the same country and often also from the same educational institution. The following are just a few examples from the investigations mentioned above: the native-speaker judges employed by Gass & Varonis (1984), Munro & Derwing (1995), and Major *et al.* (2002, 2005) were all students at particular North American universities, while those participating in the investigations by Johansson (1978), Hughes & Lascaratou (1982), and Dretzke (1985) were all linked to individual universities in Britain, or to a number of secondary schools from the same county. While the English native speakers in the first experiment described by Koster & Koet (1993) all lived in the Netherlands, those in the second experiment were all students at the University of Edinburgh. It should be noted, of course, that students drawn from a particular institution do not necessarily have the same accent; in fact, Johansson (1978: 113–114) is very careful to indicate his respondents' UK county of origin. In such studies, which employ native-speaker judges from one particular educational institution, there is very likely to be accent variation between respondents, but this is not commonly examined as a separate variable.

A single example of an experimental study of pronunciation hierarchy which distinguishes between two groups of native-speaker respondents on the basis of their accent is Johansson's (1975) perceptual study of the sounds of British English as opposed to those of Swedish. In this study, two groups of native speakers of British English (88 respondents from London as well as

25 speakers of Scottish English from Edinburgh) were asked to distinguish Swedish vowels and consonants from their nearest equivalents in RP (Received Pronunciation, alternatively termed “Standard Southern British English”).⁵ Johansson found that the Edinburgh informants were somewhat less inclined to reject particular realisations categorically as foreign, presumably because they were not as familiar with the RP model used in the experiment as were those informants with London accents (Johansson 1975: 74). However, this result may be set off against the “great similarity in overall discrimination between London and Edinburgh informants” which is “paralleled by a remarkable consistency in the reactions to individual vowels and consonants” (Johansson 1975: 75).

In those few cases where there is a salient contrast between the two groups of respondents, Johansson relates this, where possible, to differences between RP and Scottish English. For instance, the fact that the Edinburgh judges were much less inclined to reject the Swedish realisation /o:/ as an equivalent of RP /əʊ/ is attributed to “the fact that *go* in Scottish English contains a monophthong similar to the vowel in S[Swedish] *gå*” (1975: 75). Similarly, the greater tolerance attested in Scottish judges towards Swedish substitutions of /ʊ/ in *cooks* and *puss* are ascribed to the different realisations of this vowel in RP and Scottish English, and to the similarity between a Scottish centralised [ü] and a Swedish rounded, mid, central /ʉ/ (1975: 75). Johansson (1975: 75) suggests that the Scottish judges “would be less certain of the phonetic norm” for the RP vowel and would therefore “naturally more seldom reject a particular pronunciation as foreign”.

Johansson’s comparison of the perceptions of London and Edinburgh informants suggests that a hierarchy of error for English pronunciation may be affected not only by “the existence of variations in pronunciation within the English-speaking community” (Johansson 1975: 83, 1978: 93–94), but also by what Johansson (1978: 102) refers to as the “coincidental matching of the dialects of the source and target language”. The notion that judges may be influenced by similarities between non-native pronunciation features and realisations heard in their own speech community will be referred to in the present study as “accent similarity”. As Johansson (1978: 95–96) puts it: “If an ‘error’ is identical to a pronunciation which is widespread among native speakers of English, it is judged to be more acceptable”, provided that “the social prestige of different pronunciation variants [is] taken into account”. In other words, highly stigmatised realisations may be exempt from the leniency accorded by respondents to foreign realisations similar to those native pronunciations attested in their own sociolects. A similar tendency is also apparent from Swacker’s (1976) investigation of native-speaker attitudes to the use of Texas regionalisms by Jordanian learners of English. In addition, Ryan (1983: 150) has suggested that some non-native realisations of English /ð/ and /θ/ may overlap with what is heard in “lower class dialects”, which may cause non-native speakers to be downgraded socially.

⁵ For a discussion of RP, see Wells (1982: 117–120) and Roach (2004).

Despite Johansson's suggestions that these should be explored in future research, variables such as accent variation and accent similarity have not been included in any hierarchies of pronunciation error studied by the present author. Differences between, for instance, North American respondents and British respondents in evaluating the severity of non-native pronunciation errors appear not to have been studied. Since RP and GA (General American, alternatively termed Standard American English) are the most commonly taught, and aspired to, English pronunciation models worldwide, it would be useful to know how different groups of native speakers react to non-native approximations of these.⁶ In fact, no experimental studies appear to have been undertaken which attempt to establish a hierarchy of pronunciation error for any variety of American English (unless one includes studies such as Anderson-Hsieh *et al.* 1992, which discusses the relative importance attached to prosodic as opposed to segmental errors in 60 non-native speech samples by three teachers of American English). An example of a hierarchy of pronunciation error for American English based largely on observational procedures and available literature is found in Collins & Mees (1993). It would clearly be useful for teachers and students of American English, and for researchers in this area, to know how different groups of North Americans prioritise certain pronunciation problems, and whether the latter differ in this from native speakers of other varieties of English (such as RP or other British, Irish and Antipodean accents). For instance, it has been suggested by Milroy (1994: 178) that in the United States, foreign accents "seem to be more subject to negative evaluation than in Britain". This is in keeping with the view expressed by Prator (1968: 25) that Americans have a "greater antipathy toward foreign accents" (see 3.4.4 for a more detailed discussion of these matters).

1.2.4 "English as an International Language" and other approaches

Non-native learners are increasingly exposed to a wider range of standard and non-standard varieties of English in the media, and may have either instrumental or integrative motivations to model their English on that of specific speech communities where neither RP nor GA are the norm, for example those of Ireland or Australia (cf. Daniels 1995: 83). Such learners would benefit from an awareness of those characteristic pronunciation errors which are viewed either as insignificant, or as highly stigmatised, by the communities to whose English they have been exposed, and which, in some cases, may serve as their model. This does not imply that learners should adopt any realisation that is also found in native-speaker varieties, stigmatised or otherwise. For instance, Swacker (1976: 17) has argued that "[c]ertain dialectal markers may be perfectly acceptable ... when coming from a native speaker, but be quite offensive when spoken by a foreigner" (cf. Chapter 4). The use of certain substitutions for English /θ, ð/, as commented on by Ryan (1983: 150), may be a case in point, since certain of these are heavily stigmatised, especially in the US (cf. Pederson

⁶ For a discussion of GA, see Wells (1982: 118, 120–122).

2001: 260, Wolfram & Schilling-Estes 1998: 75, 161). It may therefore be unwise to suggest to learners of English that such substitutions are generally acceptable merely because they are sometimes employed by certain groups of speakers.

Nonetheless, this is one of the recommendations made by Jenkins (2000), who has attracted much recent attention through her proposal to simplify English pronunciation teaching, purportedly in the interest of non-native learners. One of Jenkins's basic assumptions is that, since non-native speakers of English greatly outnumber the natives, most communication in English actually takes place in this non-native form, which she terms "English as an International Language" (EIL). The dominance of non-native English is in fact debatable; Trudgill (2005b: 78) has argued that "there is still very much more native than non-native usage" (see also Bruthiaux 2003). However, Jenkins uses this dominance as a justification to suggest that English pronunciation training to non-native learners should therefore concentrate on teaching them to be intelligible to each other rather than to native speakers. Despite the fact that *both* goals may be achieved by teaching these learners one of a number of well-known native models of English, Jenkins (2000: 123) has instead put forward a "pedagogical core of phonetic intelligibility" for the purpose of communication in non-native English, which she has dubbed the "Lingua Franca Core".

While the imagery may be different, Jenkins's "core" is comparable to a "hierarchy of pronunciation error" in that it consists of a number of prioritised segmental and suprasegmental features that are considered crucial for a learner to acquire. As with most error hierarchies, it is essentially an attempt to "scale down the phonological task for the majority of learners" (Jenkins 2000: 123). Yet, strikingly, the most important criteria used by Jenkins are not intelligibility and acceptability by native-speaker standards, but intelligibility and learnability based on *non-native* standards. Taking up such a position entails that all aspects regarded as "unteachable" are necessarily excluded from the core. These include not only prosodic phenomena such as weak forms, but also certain segments, for example dark [ɫ] and the dental fricatives /θ, ð/ (Jenkins 2000: 138–139, 147). In this context, Jenkins makes much of the notion that these features are not found in all native varieties of English (although it is worth stating that, *contra* Jenkins, weak forms do seem to exist in all native varieties, cf. Knowles 1992: 989). The implication appears to be that if certain realisations have not been attested in all groups of native speakers, they may not be necessary for non-native communication (despite the fact that some native speakers are less intelligible to non-natives than others). Jenkins (2000: 27) presents English native-speaker variation as being generally "on a par" with variation found in non-natives – a view which is at the very least highly debatable.⁷ Nevertheless, Jenkins (2000: 139) suggests that in terms of deviation from the standard, it would be "unreasonable to have 'higher' expectations" of non-native as

⁷ In addition, it is difficult to determine which linguistic or sociolinguistic phenomena are covered by the somewhat vague term "on a par".

opposed to native speakers. Given the sociolinguistically dominant status of English native speakers over non-natives, it is questionable whether being “unreasonable” merits being a prime cause for concern.

As Jenkins (2000: 124–131) herself points out, there have been earlier pedagogically driven attempts at an English “phonological core” in studies by Gimson (1978) and Jenner (1989), but the most important innovation in the “Lingua Franca Core” is the emphasis on non-native-speaker English as a target for non-native learners. As an objective, this may appear to be somewhat paradoxical or even unnecessary, given the fact that such learners will, by definition, already have attained this target. This potential threat to the usefulness of Jenkins’s Core is redeemed by the arguably inconsistent inclusion of a number of features explicitly drawn from native-speaker models such as RP and GA, which happen to be “crucial to intelligibility among L2 ... speakers of English (Jenkins 2000: 131). As Gibbon (2005: 450) suggests, this “coincidence” is in fact very convenient to Jenkins. As things stand, it is notoriously difficult to define intelligibility (or even learnability) by the different standards of non-native speakers of English from widely diverging backgrounds, without recourse to a native model (cf. Trudgill 2005b: 80–82, 86–88, Gibbon 2005: 450; see also 3.7). In this way, the native-speaker norms enter, as it were, through the back door (cf. Gibbon 2005: 450, Scheuer 2005: 114–115). If, in fact, native-speaker norms were completely irrelevant, it should not matter whether some non-native variation is, as Jenkins (2000: 27) puts it, “on a par with that which we find among L1 accents of English”.

If the proposed non-native model is, as Jenkins (2000: 131) puts it, “grounded in RP and GA” ostensibly in order to safeguard mutual intelligibility for non-natives, one wonders why a native English model has not simply been recommended instead. As Trudgill (2005b: 92) has argued, “Jenkins’ proposal is totally equivalent to many forms of native-speaker Irish English ... and also to Standard Jamaican English”. Trudgill (2005b: 92) also points out that the “only difference is that she is happy to permit many, though not all, [non-native] phonetic realisations”, and goes on to summarise this as: “aim at Irish English but don’t try so hard as before with the phonetics”. In a separate footnote, Trudgill (2005b: 92, n.11) argues that since most learners fall short of pronunciation targets set to them (whether native or non-native), such an attempt to “relax the phonetic target for EIL” may well lead to a reduction of intelligibility.

There are at least two other reasons why Jenkins (2000) insists on a non-native model of English for EIL, which may be summarised as “learnability” and “native-speakerism” (cf. Holliday 2005: 6). Jenkins evinces a commendable concern with the ability of non-native speakers to acquire certain features of English pronunciation; she discusses a number of theoretical approaches to second-language acquisition which, she claims, account for “the complicated combination of factors involved in transfer” from learners’ native language to English (Jenkins 2000: 119). These include the Markedness Differential Hypothesis (Eckman 1977; for a more recent discussion, see Leather 1999: 30), which

seeks to predict areas of learner difficulty on the basis of relatively infrequent or “marked” realisations found in the target language. An example may be provided by the dental fricatives /θ, ð/, which are described by Jenkins (2000: 101) as a “universally difficult feature of English”. As a result of this, they are eligible for omission from the Lingua Franca Core. Their substitution by other phonemes is deemed acceptable, especially since, as Jenkins (2000: 120) claims, /θ, ð/ are “not relevant to EIL intelligibility” (see also Jenkins 2000: 134, 137).

It is of course true that dental fricatives are relatively infrequent phonemes among the languages of the world, which, at least from the point of view of learner difficulty, would at first sight appear to argue for the acceptability of their substitution by other articulations. However, such substitutions are not consistent among learners of different L1 backgrounds, which could lead to learner confusion and potential intelligibility breakdown. Furthermore, those substitutions may also be problematical to speakers of those languages which have either or both of these sounds in their consonant inventories (including Greek, Welsh, Icelandic, Arabic and Castilian Spanish). In addition, informal observation indicates that certain groups such as speakers of Latin American varieties of Spanish and some Southern African languages (e.g. Swahili) seem to have relatively little difficulty in producing convincing articulations of /θ, ð/, even though these sounds do not function as full (as opposed to marginal) phonemes in those languages (Beverly Collins, personal communication).

There are also other ways of establishing the relative significance of certain phoneme substitutions made by different groups of learners. One such method which is directly relevant to establishing a hierarchy of error is Brown’s (1988) study of functional load. Brown (1988: 215, 218, 221) provides an analysis of a number of factors which may affect the intelligibility of an utterance if a particular RP phoneme is substituted by a phonetically similar one, such as the question of how many minimal pairs are distinguished by the contrast between the two phonemes, and the cumulative frequency of the contrast in question. This implies that a contrast such as /θ ~ s/ has a much higher functional load than /θ ~ f/ (Brown 1988: 222), and therefore poses a greater threat to intelligibility than the latter. In other words, not all substitutions of dental fricatives necessarily have the same effect on intelligibility.

There have been many other discussions of learner difficulty as predicted by research into interlanguage phonology, but it would be outside the scope of this dissertation to discuss these in full. An overview is, for instance, provided in Leather (1999: 37–38). Such discussions tend to focus on the order in which certain features should be taught, or the environments in which they are most or least likely to be problematical. What Jenkins is proposing, however, is that certain features of English which are highly marked should not be taught at all – unless this creates intelligibility problems for other non-native listeners. This implies that in EIL, learner difficulty is prioritised over intelligibility and acceptability by native-speaker standards. It is questionable to what extent a learner who has been taught English by such standards is still capable of communicating efficiently with native speakers – or even with other non-native

speakers – regardless of whether the latter have modelled their English on native-speaker varieties.

Nonetheless, Jenkins (2000: 211) has described attempts at promoting non-native intelligibility for native speakers as “anachronistic” and “doomed to failure”; as she puts it, “there is no good reason to expect learners to acquire these [native-speaker] features and, by implication, in the process to obliterate as much as possible of their L1 accents and, along with these, their L1 identities”. This would almost appear to suggest that non-native learners do not continue to have ethnic and linguistic identities of their own outside communication in English – a somewhat ethnocentric, if not Anglo-centric, notion indeed. Be that as it may, Jenkins claims that it is in fact native speakers of English who now have to make themselves intelligible to the learners instead. As is announced somewhat dramatically by Jenkins (2001: 227):

The perhaps unpalatable truth for “N[ative] S[peaker]s” is that if they wish to communicate in international communication in the 21st Century, they too will have to learn EIL. For future children, it can be incorporated into the secondary school curriculum as a compulsory component of their existing English studies, and alongside the learning of other languages. (...) For those who have already reached adulthood it will be necessary to attend adult EIL classes in the same way that “N[on-] N[ative] S[peaker]” adults do.

In the unlikely event that this ever comes to pass, the need to be intelligible to native speakers, already decreed “anachronistic” by Jenkins, will indeed be seriously reduced. Similarly, there will be no need or opportunity for foreign learners to attempt to acquire a native-like pronunciation, as all native speakers will have been trained to address non-natives in EIL. Seen from this millennial vantage point, Jenkins (2000: 161) can actually afford to be generous when she states that she has “no desire to patronize those learners who wish to sound ‘native-like’ by telling them that they should/need not go to such lengths”: sooner or later, the projected reality of EIL communication will catch up with them (see also Trudgill 2005b: 94–96).

Jenkins’s predictions about the fate of the native speakers do not merely derive from her evident enthusiasm about the EIL project, but are also fuelled by ideological rather than didactic (or even purely linguistic) perceptions of the roles of natives and non-natives. A central tenet of her study is the notion, based on the following notorious claim by Widdowson (1994: 385, quoted in Jenkins 2000: 7) that the English language is not actually “owned” by its native speakers:

How English develops in the world is no business whatever of native speakers in England, the United States, or anywhere else. They have no say in the matter, no right to intervene or pass judgement. They are irrelevant. The very fact that English is an international language means that no nation can have custody over it.

Seen like this, native-speaker evaluations of non-native speech are merely expression of a “native-speakerist” bias, which assumes that, as Holliday (2005: 8) puts it, “‘native speakers’ of English have a special claim to the language itself, that is essentially their property”. It is clearly this imbalance between natives and non-natives that Jenkins’s EIL is trying to address, with the possible result that the native speakers will find themselves “on the receiving end” of pejorative stereotyping (Jenkins 2000: 219).

Implicit in Jenkins’s argument seems to be the idea that non-native speakers will feel relieved upon hearing the news of the EIL endeavour. It is dubious, however, whether all non-natives appreciate the favour. Many learners do in fact use native-speaker English as a model for their speech, or aspire to do so. Major *et al.* (2005: 44) have pointed out that non-native speakers, rather than being a kind of mutual admiration society, may also be biased against non-native English. As a result, they may hold serious objections to pronunciation models such as Jenkins’s EIL (cf. Scheuer 2005: 126–127). As Christophersen (1973: 85) has shown, when presented with “native-speaker” schemes to simplify English pronunciation designed with the interests of non-native speakers in mind, learners are not unlikely to reject “indignantly the idea that ‘normal’ English [is] to be withheld from them”. Interestingly, Holliday (2005: 164) describes the “native-speakerist” approach to TESOL (Teaching English to Speakers of Other Languages) as an attempt on the part of English-speaking Western educators to control “what sort of English people should speak”. He suggests furthermore that Jenkins (2000) may also unwittingly be engaged in this practice, despite her “intention to democratize English by arguing for an international standard”. In this context, Holliday (2005: 165) goes on to mention “Western TESOL ‘philanthropists’” caught in a “liberation trap, where the supposedly democratizing English-speaking Western TESOL discourse is not appreciated by the people it is supposed to be helping and imposes its own construction upon them”. This may be seen as a covert reference to the “Lingua Franca Core”, developed by a native-speaker teacher of English for the benefit of non-native learners.

Given the contentious nature of Jenkins’s proposals, it is to be expected that her suggestions have met with support and approval (Seidlhofer 2001, McKay 2002, Seidlhofer 2005) on the one hand, but also with largely negative or hostile reactions, from native as well as non-native researchers (see Dziubalska-Kołaczyk & Przelacka 2005), on the other. It can certainly be questioned whether teaching EIL is the best way of enabling foreign learners to deal with the native versus non-native power imbalance; lack of access to native English may cause non-native learners to be downgraded socially and professionally (cf. MacKenzie 2003: 61, Scheuer 2005: 125). In other words, legitimate concerns over native-speaker sociolinguistic dominance should not end up with proposals which, once categorically adopted by English-teaching institutions, may lead to non-natives being denied access to the native models they may require. Viewed from this perspective, it is important to establish what the effect would be of applying Jenkins’s “hierarchy of error” for EIL to

pronunciation teaching in different contexts. If it is found that non-native speech which is “acceptable” by the criteria of the Lingua Franca Core is actually downgraded by native speakers, this would present an additional reason for not introducing EIL teaching in contexts traditionally associated with EFL. In the present study, Jenkins’s explicit suggestions for her “Lingua Franca Core” will be compared and contrasted with the findings for a hierarchy of error for Dutch learners of English.

1.3 Objectives and methods of the present study

1.3.1 General and practical objectives

Despite the fact that in some quarters native speakers have been decreed “irrelevant” to learner objectives in second language acquisition, this dissertation maintains the position that large groups of L2 learners of English will continue to exist who wish to communicate effectively with native speakers (and with other non-natives who adhere to native models). Furthermore, native speakers will continue to be a valuable source of information to such learners. This is why this dissertation is emphatically intended as a contribution to research into *native-speaker* attitudes to non-native speech.

The discussion in 1.2.1 has shown that L1 assessments of accented speech are preponderantly negative, while at the same time an L2 accent is almost impossible to eliminate. This unfortunate state of affairs implies that, until native speakers are actually made to change their attitudes to accented speech (as is suggested in Jenkins 2000: 227–229), learners may wish to modify those characteristics of their speech which, by rendering their speech less intelligible or more irritating or distracting, most seriously reduce their communicative efficiency. It is one of the primary aims of this study to determine which speech features are most eligible for training.

This has been done within the context of severity judgements, by different groups of native speakers, of representative segmental and suprasegmental pronunciation errors as found in Dutch learners of the two most generally taught accent models of English worldwide: RP and GA (see 1.2.3). In the same vein as earlier studies in this area (see 1.2.2), attempts were made to rank these errors by significance in an overview called a hierarchy of error. These results, based on native-speaker judgements, may then be compared with the recommendations made by Jenkins (2000) for the benefit of non-native communication, so that the differences and similarities between the effects of the two approaches on error prioritisation are clearly visible. The results may also be contrasted with other error hierarchies that are relevant for the Dutch situation (see below).

Where this study differs from all previous attempts to establish such a shortlist of significant pronunciation problems for learners of English (see 1.2.2 and 1.2.3) is in the exclusive employment of empirical evidence collected from

different groups of L1 judges to establish two separate hierarchies: one for RP and one for GA. This considerably widens the scope of research into error prioritisation. For instance, it makes it possible to determine if pronunciation errors are perceived and ranked differently in different standard varieties of the same language.

The effects of rater variables on hierarchy of error have not previously been exhaustively studied (see 1.2.3). Nevertheless, there are some indications that age, sex, attitude and linguistic sophistication play a part in this. More attention has been given to comparisons of native versus non-native raters, and instructors versus non-instructors. Most of these variables are also considered in the present study, but a special emphasis has been placed on the effect of native judges' language background. It has already been noted by several workers in the field that the specific L1 variety spoken by respondents may affect their judgements of non-native speech (e.g. Johansson 1975, 1978, Politzer 1978, Piazza 1980). In addition, it has been suggested that particular native-speaker groups, such as Americans, may rate L2 speech more strictly than do, for instance, British judges (e.g. Prator 1968, Milroy 1994). However, as pointed out in 1.1, it is the belief of most Dutch learners that the British and Irish evaluate non-natives more strictly. Such stereotypical views could have a considerable effect on learner perceptions and motivations.

The above suggestions lead one to conclude that "native-speaker accent" could well be a significant rater variable to be studied in the context of hierarchy of error. Consequently, it is an important objective of the present study to investigate the effect of native-speaker accent on judgements of non-native speech. For reasons of expediency, the investigation has largely focused on regionally distinctive accents, as opposed to accent variation related to other factors, such as class or ethnicity. In spite of this limitation, such a line of inquiry may well prove to be a useful contribution to further research in this area.

It has been suggested that non-native pronunciation features which appear to have equivalents in native speech will be evaluated less severely by judges from the relevant speech community. Very little research has been carried out on this subject, but this assumption has, for instance, been made by Johansson (1978: 95–96), and it also appears to be a principle that is observed in Jenkins's *Lingua Franca Core* (see 1.2.4). Within the context of pronunciation training, there is anecdotal evidence (personally confirmed by the present author in pronunciation classes given to Dutch students of English at two different universities in the Netherlands) that advanced and linguistically sophisticated learners explain their incidental deviations from an L2 pronunciation model not as L1 interference, but as an attempt to imitate regionally distinct features – suggesting, for instance, that their realisation of /ð/ as a stop rather than a fricative, which is a characteristic feature of Dutch English, would be acceptable in Irish English.

It is another core objective of the present study to test the hypothesis that accent similarity positively affects native-speaker reactions to non-native

speech. If this is indeed the case, it may also be used as an explanation to account for differences between groups of native-speaker judges. It may be, however, that certain realisations are so stigmatised as to counteract the effect of accent similarity (cf. Ryan 1983, Johansson 1978). Be that as it may, both possibilities suggest that native-speaker judges sometimes evaluate certain specifically foreign pronunciation features not simply as L2 speech, but also by the standards of what they consider to be acceptable L1 regional or social variation. This arguably provides a new dimension to investigations into native-speaker attitudes to foreign-accented speech. It also addresses the topical questions of to what extent regional or social accents are to be recommended to non-native learners as target models (cf. Daniels 1995: 83), and to what extent learners should be encouraged to use specific regional features in their speech. This is especially relevant to pronunciation training at an advanced level.

It may be clear that the objectives of this dissertation, as described in the above, all reflect a desire to place studies of hierarchy of error firmly in a variationist context. This is partly motivated by the dearth of sociolinguistically oriented research in this area, but also by an awareness of English not as a monolith but as a pluricentric language, with different norms being associated with the different standard varieties (cf. Clyne 1992). This awareness is not only increasingly evident in English-language research, but also in learners' attitudes to the different varieties (e.g. Van der Haagen 1998, Ladegaard 1998, Preisler 1999). As was pointed out in 1.1, learners' exposure to the different accents of English, as facilitated by the modern media, has also dramatically multiplied the opportunities for confusion. Globalisation has also made it far more urgent to prepare learners for interaction with native and non-native speakers with widely divergent linguistic and cultural backgrounds. All this would suggest that analyses of hierarchy of error from a variationist point of view would also be beneficial to research into second language acquisition, and its practical applications in pronunciation training.⁸

On a practical level, the status of English as a pluricentric world language means that it may be important for learners (especially at an advanced level) to be aware of the pronunciation norms of different speech communities – including those associated with the target accent they may be motivated to imitate, whether this is RP, GA or some other accent. Such information may be made available to learners as part of their pronunciation training, and could be incorporated into pronunciation manuals aimed at EFL students. Establishing

⁸ Such analyses could theoretically include investigations of how particular non-native pronunciation features are received by other specific groups of non-native learners. As Setter & Jenkins (2005: 3) suggest: "If intelligibility between [native speakers and non-native speakers] is a source of data for researchers, intelligibility in English between [non-native speaker] groups would seem to provide endless possibilities for research, and could lead to the development of teaching materials which are geared towards particular English communication situations – between Hong Kong and Japanese speakers of English, perhaps. The scope for study, then, is almost infinite".

hierarchies of error for RP and GA, and comparing and contrasting the various priorities given to learner errors by different groups of native speakers, will be a step towards achieving this goal.

Needless to say, the most effective way of making learners aware of such different speech norms would be to discuss them in relation to their own L2 accents in English. Accordingly, the present study has employed realistic pronunciation problems which have been attested in the actual training of advanced Dutch learners of standard British and American English. If, on this basis, error hierarchies can be established for two influential varieties of English, these data can be directly implemented in pronunciation training in the Netherlands, in the form of pronunciation manuals and other such textbooks. Additionally, the focus on native-speaker accent as a variable will help to provide more information on how speakers of different varieties of English (including not only Australians, Canadians, Irish people, New Zealanders and South Africans, but also the very large numbers of Americans and Britons who do not speak either RP or GA) evaluate features of a Dutch accent in English. This knowledge will be directly relevant to those Dutch learners who are motivated to model their accents on these varieties. Similarly, it would be very helpful, both to teachers and learners of English, to be aware of any mitigating effects of accent similarity on characteristically Dutch pronunciation errors in English.

As a result of new insights of this sort, it may be possible to realign priorities in pronunciation teaching in the Netherlands. This is an important practical goal of this dissertation. To this end, the results of the present study were compared with those of existing hierarchies of error that are relevant for the Dutch situation. This meant that the hierarchy for RP has been compared and contrasted with similar studies by Collins & Mees (2003b: 290–293), and in its modified form in Collins *et al.* (1987), by Gussenhoven & Broeders (1997: 16–17), and by Koster & Koet (1993). Similarly, the hierarchy for GA has been subject to comparison with a similar one drawn up by Collins & Mees (1993). It has also proved interesting to compare the results of this dissertation to those found by Dretzke (1985), who conducted a very similar experiment in order to establish a hierarchy of error for German learners of RP. The findings of the present study as regards error prioritisation by native speakers have also been held up for comparison against two very different approaches, namely Jenkins's (2000) *Lingua Franca Core*, and Brown's (1988) study of functional load (see 1.2.4). In this way, the relevance of these studies and approaches with regard to the Dutch situation can be subjected to appropriate scrutiny.

In order to ensure that the present study would meet its objectives, it was empirically grounded in the actual practice of teaching pronunciation to Dutch learners. The pronunciation problems presented to the native speakers were not only drawn from pronunciation manuals primarily aimed at the Dutch market, but were also selected by five groups of Dutch judges prior to their inclusion. These groups consisted mostly of non-native teachers and students of English, and university and college lecturers employed in departments of English in the

Netherlands. As it was actually their selection of errors that was subsequently presented to the native-speaker judges for evaluation, it was assumed that this would make it possible to compare and contrast the Dutch judges' error prioritisations with those of the native-speaker judges. Such a comparison is in the same vein as other studies of hierarchy of error which have investigated differences between L1 and L2 judges (e.g. Johansson 1978, Galloway 1980, Hughes & Lascaratou 1982, Sheorey 1986, Fayer & Krasinski 1987, Schairer 1992, Koster & Koet 1993, Rifkin 1995). If any disparities are found between the native speakers and the Dutch judges, this may well prove to be valuable information which could be used to realign priorities in Dutch pronunciation training in the Netherlands. The same purpose would be served by the questions addressed at the Dutch judges with regard to their attitudes to pronunciation training in the Netherlands. These questions are discussed in detail in 2.1.2; an analysis of the Dutch participants' responses is provided in 2.2. The latter serve as an indication of the relative importance attached to English pronunciation training in secondary and tertiary education in the Netherlands.

1.3.2 Overview of the methods used

This dissertation has employed two experiments in order to achieve the objectives as stated in 1.3.1. For the sake of convenience, these have been labelled "the Dutch Experiment" and "the Native-speaker Experiment", since judges participating in the former were all resident in the Netherlands and the survey was conducted in Dutch, whereas those taking part in the latter were all self-identified native speakers of different varieties of English. As the Dutch Experiment chiefly served to select a number of representative test items which could be used in the Native-speaker Experiment, the latter will also be referred to as the "core experiment". A detailed account of the methods used in these experiments is provided in 2.1 (for the Dutch Experiment) and 2.4 (for the Native-speaker Experiment), but an introductory overview will also be provided here.

The experiments were primarily constructed in order to be able to establish an error prioritisation for Dutch learners of RP and GA, and to study the effect of different variables, especially linguistic background, on respondents' judgements of the severity of these errors. To begin with, the Dutch Experiment was designed to collect evaluations from non-native respondents who were sufficiently experienced with second language acquisition in English to rate a number of representative pronunciation errors on the basis of verbal descriptions. Respondents rated the errors on what is termed a Likert scale (Likert 1932) ranging from "no error" to "a very serious error". The design of the Dutch Experiment also permitted the collection of data on respondents' age, sex, linguistic and professional/educational background, and attitudes to pronunciation training. In order for the errors to be representative, they had to be largely drawn from an existing corpus (namely the hierarchy of error provided by Collins & Mees 2003b: 290–293) and belong to a number of discrete categories (see 2.1.3 for detail). For each of these categories, those errors identified as

being among the most serious by any group of Dutch judges served as a basis for the items presented to the native-speaker respondents in the core experiment.

Similarly to the Dutch Experiment, the Native-speaker Experiment was also designed to allow different groups of judges to rate a selected number of pronunciation errors on a Likert scale with the same ranges of severity. Since the selection of the errors was based on that provided by the Dutch respondents, this made the test items more representative. Moreover, it enabled one to compare respondents' assessments of particular errors in the two experiments. This allows a comparison of the effects of the variable "native versus non-native". Most importantly, the error ratings thus obtained permit a hierarchy of error to be established on the basis of native-speaker judgements. As in the Dutch Experiment, no attempt was made to define error gravity in terms of different potential effects (such as "unintelligible" or "distracting"). Apart from the fact that it is questionable whether respondents can reliably distinguish between these effects (see 1.2.2), such further specifications could bias judges against particular types of error, and might not accurately reflect their reasons for assessing their gravity. In addition, the design of the core experiment also allowed for data collection on respondents' age, sex and linguistic background.

Despite the similarities between the two experiments, they served different purposes – the Dutch Experiment being intended primarily to help select errors for the Native-speaker Experiment. This meant that there would also be differences in design and in the type of data gathered. For instance, while it was considered important to collect information on the professional background of the Dutch judges, since this could reflect on their attitudes to pronunciation training in the Netherlands, no similar data were collected in the native-speaker survey, as an investigation of this variable was not a direct aim of this experiment. In fact, since native speakers' linguistic sophistication may be a hindrance in research of this nature (cf. Johansson 1978: 22), it was decided not to factor respondents' educational or professional background into the design of the core experiment. It was, however, considered useful to ask native-speaker participants to describe their own leniency as judges of pronunciation (see 3.1.2). In addition to being indicative of respondents' general attitude to pronunciation errors, such information about leniency makes it possible to normalise judges' assessments for this particular variable. Since one of the main aims of this dissertation was to study the effects of accent variation on native-speaker judgements, all respondents in the core experiment were also asked to describe their own accent in English. These self-identifications were then categorised by the researcher on the basis of the available literature on accent variation in English. On the basis of these categories, participants were divided into a number of accent groups, which could then be employed as a separate variable.

It was also seen to be in the interest of representativeness that the core experiment would be directed at various different groups of native speakers, and should therefore not presuppose any familiarity either with Dutch or with linguistic terminology. This implied that instead of the verbal descriptions used

in the Dutch Experiment, native-speaker judges would be presented with recordings which feature the error in an otherwise authentically native-like context. This would involve the use of phonetically trained bilingual actors capable of making a single Dutch pronunciation error, while delivering the rest of the utterance in a convincing native English accent. Respondents could then be asked to identify the single error and evaluate its severity. A similar technique had previously been used in Johansson's (1978) study of hierarchy of error, in which one of the investigations featured a bilingual speaker of Swedish and RP-accented English who, as Johansson (1978: 89) put it, had been asked "to 'fake' particular kinds of pronunciation errors in the sentences" concerned. Similarly, Dretzke (1985: 93) employed a bilingual speaker of near-RP and German to the same effect, noting that this was essentially a variation on the "matched guise" technique (cf. Lambert 1967), in which the impression of multiple accents is created by one speaker, so that attitudes to these accents can be measured while control is maintained over all other variables. However, one difference between traditional matched guise (Giles 1970, 1971) on the one hand, and both Dretzke's investigation and the present study on the other, is that while the actors' accents are kept constant, the errors vary in each utterance presented to the respondents.

Given the fact that the core experiment was partly designed to study the effect of accent variation, it was considered that more than accent guise should be made available to the judges, and that the latter would be allowed to choose the accent that they felt most competent to judge. In view of the objectives stated in 1.3.1, but also as a result of the availability of different actors, it was decided to introduce only two guises: one for RP and one for GA. Whilst this clearly does not cover the range of standard accents that learners of English could conceivably select as a model, it does at least make it possible for native speakers of accents other than these to refer to one of two internationally recognised varieties of English when evaluating the pronunciation errors included in the experiment. (One of the experiments conducted by Johansson indicated that respondents from Edinburgh were quite capable of judging Swedish pronunciation errors in RP-accented English.) This also permits comparison of error prioritisations judged by the standards of GA as opposed to those for RP, provided the two guises are sufficiently similar in all other aspects. Most importantly, it has facilitated the construction of two separate hierarchies of error for the two most widely taught accent models of English – a fundamental aim of this investigation.

In order to be able to investigate the variable of accent variation, it is of course not sufficient merely to provide different accent guises. In addition, a wide variety of differently accented speakers of English have to be able to participate in the experiment, so as to make the selection of respondents as divergent and representative as possible. Accordingly, it was decided to present the core experiment in the form of an Internet survey, which rendered it possible for volunteers from all over the world to take part in the experiment without being tied to a particular location. This implied (1) that the audio stimuli would

have to be provided as easily downloadable sound files, (2) that respondents would have to be able to detect and assess the relevant errors online, and (3) that their assessments would be made accessible to the researcher. A special software program was developed to deal with these requirements. (The format of a web survey was also used for the Dutch Experiment, but this did not require any of the special features necessary for the core experiment.)

In the course of conducting these investigations, it became apparent that, as a result of the differences between the two experiments, it would be difficult to compare and contrast the different error evaluations of the native and non-native judges reliably in all aspects. Even though this was only a subsidiary objective of the present study, it would still be useful to be able to make this comparison. Accordingly, an attempt was made to compare the differences between the Dutch and native-speaker judges using regression analysis. A discussion of the difficulties involved in such a comparison, and of the method used, is provided in 3.6.

Similarly, it should be pointed out that when the two experiments were originally designed, the notion of accent similarity as a mitigating factor on judges' assessments of error gravity had not been taken into consideration. It was only as a result of accumulating insight on the part of the researcher that it was decided to determine the possible effect of accent similarity, and to elevate this notion to being one of the central concerns of the present study. This meant that, after the errors had already been selected for inclusion in the core experiment, the researcher proceeded to make an inventory of those errors which, as a result of their similarity to native-speaker realisations, would be eligible for inclusion in a post-hoc analysis. Based on a wide variety of handbooks, and studies dealing with dialectology and accent variation, it was found that no fewer than 20 errors were similar to realisations associated with one or more of the 22 accent groups represented in the core experiment. In a few cases, such realisations had been attested in a majority of speakers of the accent concerned. Most, however, were only found in a minority of speakers of the different varieties studied. For the other 16 errors included in the core experiment, no such correspondences were found. It was on these grounds that the hypothesis was tested that accent similarity would positively affect judges' evaluations of error severity (for more detail, see Chapter 4).

1.3.3 A note on statistics

In both experiments, judges were asked to rate the severity of a number of pronunciation errors on a 5-point Likert scale. Such ratings are subject to two sources of random variation, i.e. (1) between judges and (2) within judges between items. Traditional statistical techniques cannot discriminate between these two types of variants, which leads to the employment of inappropriate statistical models (Snijders & Bosker 1999). For example, if each judgement were regarded as an independent observation, then the obvious correlation among multiple responses from one judge would be ignored. This is the so-called "design effect", which may lead to an increased chance of a Type I error,

i.e. the incorrect rejection of the null hypothesis (Quené & Van den Bergh 2004: 106). Similarly, if responses were averaged for each judge, then one would ignore the within-judge variance in responses, and decrease the number of observations, thus reducing the statistical power of the study in question (Quené & Van den Bergh 2004: 118). Such “aggregation” (Snijders & Bosker 1999: 14) would prevent us from studying the interaction, in the present study, between a judge’s linguistic or educational background and his or her pattern of responses.

What is needed, therefore, is an analysis which can take into account both types of variance simultaneously. The present study employs multi-level analysis for this purpose (Kreft & De Leeuw 1998, Luke 2004, Quené & Van den Bergh 2004). In the experiments discussed in this study, the response data will be regressed on several fixed predictors such as linguistic or educational background, sex, age, etc., while taking into account both variance between judges (level 2, i.e. higher level) and between items, nested within judges (level 1, i.e. lower level). This means that the resulting regression coefficients are “corrected” for the random variation between and within judges. This is the first study to investigate the interaction between different groups of native-speaker judges and the items they judge, by means of this statistical technique. The multi-level analyses in this study were all performed with the MLwiN program (Rasbash *et al.* 2000); this program allows for great flexibility in specifying the model, even though it is not user-friendly for beginners (cf. Quené & Van den Bergh 2004: 119).

CHAPTER 2

DESIGN AND SET-UP OF THE TWO EXPERIMENTS

2.1 The Dutch Experiment: design, subjects and procedure

2.1.1 General aims and target groups

The Dutch Experiment had two general aims. It was designed to elicit (1) severity evaluations of a number of well-known Dutch pronunciation errors in L2 English, and (2) views on the role of English pronunciation teaching in secondary and tertiary education in Holland, from groups of respondents in the Netherlands that appear to be well-placed to provide these. A large majority of Dutch secondary school teachers and university students of English, and university and college lecturers in English language and literature, will either have been taught English pronunciation or will have taught the subject themselves. This is why it was assumed that they would be sufficiently familiar with the subject, be linguistically sophisticated enough, and also have adequate knowledge of educational practice, to provide such views and evaluations. Elements of pronunciation training feature on the curriculum of most English language and literature degree courses in the Netherlands, both at universities and teacher training colleges. Much less attention appears to be being paid to pronunciation in secondary schools, but this does not necessarily reflect on teachers' training in this subject, or familiarity with it. Rather, it may be seen as a consequence of choices made in the curriculum, in which different skills such as fluency and reading are given priority.

At any rate, teachers' views on pronunciation teaching, and any possible priorities when it comes to error evaluation, are highly relevant. Since English is a compulsory school subject in the Netherlands, some of their opinions and attitudes are likely to have percolated down to the general public, and may continue to affect future generations. It would be useful to establish, by means of the present experiment, which types of pronunciation errors are prioritised most by teachers of English in secondary schools, by university lecturers and also by students of English, dependent on their assessment of the urgency of the errors in a pedagogical context. If these priorities were compared and contrasted with native-speaker evaluations of the same errors, this could assist in the realignment of priorities in pronunciation training in the Netherlands. Such a comparison would also help to determine whether or not non-native speakers, and teachers in particular, judge pronunciation errors differently from native speakers, both with regard to overall severity and in terms of attention to particular types of error. It has, for instance, been argued (see 3.6 for details) that native speakers give more priority to "global" or "prosodic" errors than do

non-native speakers (see Ellis 1994: 66, Johansson 1978: 9–15, 123), and that the latter group may be more fastidious, or less capable of reliably detecting L2 accents (see Koster & Koet 1993: 69, Bongaerts 1999a: 9). It is one of the goals of this study to assess the validity of such arguments.

The first experiment was originally set up as a paper-and-pencil questionnaire consisting of four sections (discussed below), presented in Dutch, and aimed at secondary school teachers of English. Letters were sent to different types of secondary school in Utrecht, Leiden and surrounding areas, either addressed to the school managers or to members of the English section, asking for their help in participating in a survey about English oral proficiency. This was followed up, where possible, by a phone call. While this resulted in only a few positive reactions, as a result of which the researcher went to the schools concerned and waited while the teachers filled in the questionnaire, there were also a surprisingly large number of refusals and evasions. At one school in Breukelen (near Utrecht), the head of the English section even refused any further cooperation on discovering that the questionnaire was concerned exclusively with pronunciation. As he pointed out in a forceful letter afterwards, he felt it to be misleading that investigations into pronunciation were presented as research on oral proficiency. Further objections were made to the survey's emphasis on phonetics, which, as a scientific discipline, he described as being completely irrelevant to teachers. Whilst this might increase any researcher's determination to discover if such resistance to phonetics-driven pronunciation teaching as an integral part of oral proficiency was more widely shared, it also suggested that response may be improved by enabling potential participants to see the actual questions before agreeing to take part.

Both to increase response, and to save time and expense, it was decided to revamp the paper-and-pencil survey as an online questionnaire.¹ An Internet link to this questionnaire would be placed in e-mails directed at English teachers in various types of schools in the Netherlands, ranging from pre-vocational training to pre-university education, regardless of which years they taught. In addition, teachers would be offered the chance to take part in a lottery for a number of book tokens if they submitted their e-mail address (so that they could be contacted about any prizes). However, they could also participate anonymously. This approach resulted in no fewer than 101 submissions, 98 of which were found to be usable. (In three cases, individuals had produced multiple submissions.) These include the results of the earlier paper-and-pencil surveys, the original data of which were resubmitted by a research assistant using the online format.

¹ The survey was accessible online from February to April 2001 from a personal website created by the researcher within the domain of the Faculty of Arts of the University of Utrecht at www.let.uu.nl/~rias.vandendoel/personal/enquete_html2.htm. The results for each survey were sent as anonymous e-mails to the researcher's university e-mail address.

It may perhaps be argued that this sample may not necessarily be representative because only volunteers able and willing to complete electronic questionnaires will have taken part, and unsolicited response has not been fully excluded. According to Clayton (2004), these are common issues in web-based surveys. Similar reservations may be voiced about other versions of this electronic questionnaire, or about the second online experiment. However, if these effects are at all relevant in this context, especially since Internet access is widespread in the Netherlands, and the medium is used very actively in educational contexts, they may have been compensated by the relatively large sample size, which may serve to decrease the sampling error – as Clayton (2004) also suggests. Four other versions of the same online survey were also prepared, with some of the questions slightly altered so as to be relevant to the other groups of respondents to be targeted (see below). The texts of the accompanying e-mails were also changed accordingly. Two of these versions were aimed at Dutch university students of English: one at secondary school pupils visiting Utrecht University as prospective students of English, and the other at students enrolled in English degree courses at the Universities of Leiden and Utrecht. The latter groups had all had some experience of articulatory and contrastive phonetics, and English pronunciation training. These students were also offered the chance to take part in a lottery for book tokens if they submitted their e-mail addresses, but again this was not obligatory. The text of the two versions was identical.² While the version aimed at the pupils elicited only five responses, the students' version generated no fewer than 96. As will be shown in 2.3, if a pairwise t-test is applied to the error severity assessments for these two versions, this reveals that it is reasonable to treat these evaluations as those of one single group (in spite of the fact that none of the pupils were likely to have undergone any pronunciation training at an academic level).

The two other versions of the Dutch Experiment were directed at (1) university lecturers in English language and literature in the Netherlands and (2) lecturers at Dutch teacher training colleges and other institutions of higher education comprising what is in the Netherlands collectively termed *Hoger Beroepsonderwijs* (generally abbreviated to *HBO*).³ These versions were only marginally different from each other. All concerned were offered the chance to take part in a lottery for book tokens provided they supplied their e-mail

² The first version was accessible online in March 2001 from another personal website created by the researcher (www.let.uu.nl/~rias.vandendoel/personal/enquetevoorlichting.htm), while the second version was available from March to April 2001 from www.let.uu.nl/~rias.vandendoel/personal/enquetestudents.htm. Again, the results were sent as anonymous e-mails to the researcher's university e-mail address.

³ In the glossary provided by NUFFIC (Netherlands Organisation for International Cooperation in Higher Education), these are referred to as “universities of professional education”. Other English translations of specifically Dutch educational terms have also been taken from this glossary (see www.nuffic.nl),

addresses.⁴ There was an exceptionally high response of 52 for the university lecturers, many of whom were from the English departments of the Universities of Utrecht and Leiden (the two universities where the researcher was employed at the time). There were also a number of responses from the English departments of the Universities of Amsterdam, Groningen and Nijmegen as well as from the Free University of Amsterdam. Conversely, there were only ten responses from the “HBO” lecturers. Statistical analysis in 2.3 will show that the severity evaluations of these two groups of lecturers may also be grouped together.

2.1.2 Sections included in the survey

Each of the five versions of the survey consisted of four nearly identical parts, the first of which contained questions to establish indexical data (age group and sex), linguistic background (native language and degree of exposure to native-speaker English) and professional/educational background. To encourage native speakers of Dutch to take part in the survey, the text of all five versions had been presented in Dutch. Since, apart from a few open-ended questions, this section was presented in a multiple-choice format, the possible answers were altered in each version to be relevant for the target group in question. For instance, students were not asked about their teaching experience but whether they had attended a Dutch secondary school. It was felt that, in view of their more limited opportunities to engage in native-speaker contact, students should only answer the question concerning whether they had spent time in an English-speaking country. Conversely, there were two additional questions for teachers and lecturers about their everyday use of English outside professional contexts, and their contacts with native speakers. In addition, secondary school teachers were requested to indicate if they taught at the level of the “basic curriculum” (*basisvorming*) and/or at the level of the “upper secondary phase” (*tweede fase*). For the university lecturers, this was replaced by a choice between phonetics, literature, proficiency, other subjects, or a combination of these, and, in the case of HBO lecturers, between the “first level” (*eerstegraads*) or “second level” (*tweedegraads*) of teacher training, or both. (As emerged from the respondents’ feedback, this involved the incorrect presupposition that all HBO lecturers are teacher trainers.) The task of formulating appropriate multi-choice responses to questions about teaching experience was compounded by the current tendency for terms in the Dutch educational system to have a relatively short “shelf life”.

The second part of the survey was identical in all versions, and consisted of a description of 40 possible pronunciation errors (including a number of distractors added as controls) which respondents were asked to rate on a five-

⁴ The survey intended for university lecturers could be accessed from March to May 2001 from www.let.uu.nl/~rias.vandendoel/personal/enquetedocuni.htm, and the survey targeted at HBO lecturers was available during the same period at www.let.uu.nl/~rias.vandendoel/personal/enqueteHBO.htm. Again, all results were sent anonymously to the researcher’s university e-mail address.

point Likert scale (see Likert 1932). The rating scale used was similar to what is known as a “semantic differential scale” but instead of using bi-polar adjectives at each end of the scale, the value of the ratings, from 0 (= no error) to 4 (= a very serious error), was provided in the instructions at the beginning of this section. This was done partly with a view to designing orderly and attractive web pages. As one of the main purposes of the Dutch Experiment was to pre-select errors for the benefit of the Native-speaker Experiment, and since most Dutch respondents were likely at least to have some familiarity with phonetics and/or pronunciation training, it was considered to be adequate to provide only verbal descriptions of the errors in Dutch, using as few phonetic terms as possible and followed by an example of the context (word or phrase) in which the errors are likely to occur. For instance, the error of /æ ~ e/ conflation was described as “rhyming *Annie* with *penny*” (*Annie laten rijmen met penny*).

The errors used in the survey were mainly selected from the error analysis provided for Dutch learners of RP English in Collins & Mees (2003b: 285–293). This is a well-known textbook in the Netherlands on the phonetics of English and Dutch, intended for both advanced Dutch-speaking learners of English and English native speakers interested in the pronunciation of Dutch. In addition to providing a contrastive description of RP and the standard accents of Dutch as spoken in the Netherlands and Belgium, it also presents a detailed analysis of the most common pronunciation errors made by Dutch learners, with an assessment of both their significance and their persistence in the form of a hierarchy of error (see Collins & Mees 2003b: vii, 285–293). A number of additional distractors were inspired by pronunciation errors noted by the researcher during pronunciation training sessions given to students of English at the Universities of Leiden and Utrecht as part of their proficiency programmes. All of these 40 errors and distractors will be discussed in 2.1.3 below. The second section of the survey was concluded by an open-ended question asking respondents if there were other Dutch pronunciation errors which they considered to be among the most significant or frequent.

The third section consisted of a number of questions about the importance attached to pronunciation training and, where applicable, the frequency, method and content of any such training as provided by the respondents. Respondents were also asked how often they actually taught through the medium of English and whether they referred to a particular native-speaker pronunciation model (such as RP or GA). Since it was assumed that the students had not had any teaching experience themselves, they were asked about their own experiences as secondary school pupils with pronunciation of English. It was hoped that this would broaden the perspective provided by secondary school teachers on this subject. Students were also asked if their secondary school teachers spoke English with a clearly recognisable accent, and whether this was British, American or Dutch. Teachers and lecturers had not been invited to evaluate their own accents, because it was assumed that this would either alienate them or cause them to describe their own accents imprecisely. This part of the survey generated a wealth of data, but it would be quite impossible to do justice to this

within the scope of the present dissertation. However, a few striking results will be discussed in 2.2.

In the fourth part of the survey, all respondents were invited to make any suggestions which they felt were relevant to research into pronunciation teaching but had not been covered by the present survey, a textbox having been provided for this purpose. This option had been exercised by 53% of respondents, notably by as many as 61% of the secondary school teachers. No clear single pattern emerges from these comments, but it is worthwhile noting that whereas some teachers stated that oral proficiency was an altogether different subject from pronunciation, others pointed out that the latter was often only taught in conjunction with the former. Any other relevant findings will also be discussed in 2.2.

2.1.3 Errors and distractors included in the survey

All 40 potential errors were presented to the respondents in each of the various Dutch versions in an alphabetical (and therefore effectively random) order, without any further indication as to the nature of the error. Many of these will be examined in considerable detail in Chapters 3 and 4, but it may be helpful to provide a short description of them here, with particular emphasis on those not included in the Native-speaker Experiment (and therefore not discussed in subsequent chapters). For this overview, it may be convenient to group these errors into a number of categories, such as **phonemic**, **realisational**, **distributional**, **stress** and **suprasegmental**. The first three terms are based on discussions of synchronic variation in native-speaker English as found in influential studies by Wells (1970, 1982: 72–80) and, in the Dutch context, by Collins & Mees (2003b: 295–296), and are here used with reference to segmental pronunciation errors in *non*-native speech.⁵ This is done in an attempt to describe the different effects that segmental errors made by Dutch learners of English may have on native speakers of English.

The term “phonemic error” refers to those L2 realisations of a particular sound that are perceived by native speakers of English as different phonemes, an example being the Dutch realisation of /w/ as [v], which is perceived by native speakers of English as /v/ (Collins *et al.* 2001: 26). While Wells (1982: 78) mentions such “phonemic oppositions” in the context of L1 “systemic variation”, as do Collins & Mees (2003b: 295), Johansson (1978: 92), for instance, also refers to “phonemic” errors in the context of L2 English, which he distinguishes from “sub-phonemic” errors (termed “non-phonemic” in Prator & Robinett 1985: xxi). Whereas these “sub-phonemic” errors are not perceived by native speakers as causing the substitution of one phoneme by another, they can still be a source of distraction. They may be divided into “realisational errors” and “distributional” errors. Realisational errors, such as the use of Dutch uvular-**r** in L2 English, involve pronunciations which are likely to be perceived by native

⁵ Note that Wells’s analysis was preceded by the seminal work of Trubetzkoy (1931) and Weinreich (1954).

speakers as unusual, stigmatised or deviant allophonic realisations of a particular phoneme (cf. Wells 1982: 73, Collins & Mees 2003b: 296). However, when learners make “distributional errors”, for instance when adopting a rhotic pronunciation while their target L1 accent is non-rhotic, they may be perceived as adding or deleting a particular phoneme. In such cases, their L2 realisations are possibly associated with the dissimilar distribution patterns of the phoneme in question in other varieties of English (cf. Wells 1982: 75–76 on “phonotactic distribution” and Collins & Mees 2003b: 296).

Admittedly, there are cases when the distinction between “distributional” and “realisational” errors may appear to be somewhat arbitrary. For instance, schwa epenthesis in the coda cluster of *film* has here been classified as a distributional error as it involves the addition of the phoneme /ə/. Arguably, it could also be categorised as a “realisational error” if [l³m] is treated as an allophonic realisation of the cluster /lm/. In addition, Wells (1982: 78) and Collins & Mees (2003b: 295) use the term “lexical-incidental” to refer to those cases “where the phoneme chosen for a word or a specific set of words is different in one accent as compared with another” (Collins & Mees 2003b: 295). It should be noted that some of the phonemic errors in the Dutch Experiment may also be described as “lexical-incidental”. For instance, the Dutch pronunciation of *colour* identically with *collar* probably originated as a spelling pronunciation (see Collins & Mees 2003b: 95). However, the error’s salience must be ascribed largely to the resulting neutralisation of the phoneme contrast between /ʌ ~ ɒ/.⁶ Such errors have therefore been categorised as “phonemic”.

Since realisations such as schwa epenthesis in *film* are also heard in some varieties of native-speaker English (see below), judges may not be prepared to classify these as errors in the context of non-native speech. This implies that all realisations which, for the sake of brevity, are described as “errors” in these and following sections and chapters should really be regarded as *potential errors*. The latter term will in fact be used occasionally where it is useful to remind readers that the status of a particular realisation as an “error” is by no means a foregone conclusion. This is also relevant in those cases where a particular realisation may serve as distractors because it is unlikely to be regarded as an “error” in one or more varieties of English. It should also be noted that the term *intended error* has been reserved to refer to those which the researcher wished to include in one of the experiments described in this dissertation, and to distinguish them from those cases where respondents objected to a different error, or a different aspect of the same error, from that intended by the researcher.

Table 2.1 presents an overview of all the potential **phonemic** errors described in Dutch in the relevant experiment (with English translations and keywords in SMALL CAPITALS to facilitate easy reference).

⁶ This has been immortalised in the Dutch television commercials for the detergent “OMO COLOR”, broadcast over many years, where the second part of the product’s name is pronounced as [kəʊlɔ̃r].

Table 2.1. Descriptions in Dutch of the **phonemic** errors as presented to participants in the Dutch Experiment, arranged in alphabetical order according to keywords, and with English translations.

Key word	Dutch description of error	English translation
ANNIE	<i>Annie</i> laten rijmen met <i>penny</i>	rhyming <i>Annie</i> with <i>penny</i>
BAT	geen verschil tussen de klinkers in <i>bet</i> en <i>bat</i>	no distinction between the vowels in <i>bet</i> and <i>bat</i>
BED	geen verschil tussen de laatste medeklinkers in <i>bed</i> en <i>bet</i>	no distinction between the final consonants in <i>bed</i> and <i>bet</i>
COLOUR	geen verschil tussen de klinkers in <i>collar</i> en <i>colour</i>	no distinction between the vowels in <i>collar</i> and <i>colour</i>
EXAM	<i>exam</i> laten rijmen met <i>jam</i>	rhyming <i>exam</i> with <i>jam</i>
OFF	geen verschil tussen de medeklinkers in <i>off</i> en <i>of</i>	no distinction between the consonants in <i>off</i> and <i>of</i>
PULL	geen verschil maken tussen <i>pull</i> en <i>pool</i>	making no distinction between <i>pull</i> and <i>pool</i>
SURE	geen verschil tussen <i>sure</i> en <i>shore</i>	no distinction between <i>sure</i> and <i>shore</i>
THAT	geen juiste th in het woord <i>that</i>	lack of appropriate th in <i>that</i>
THIN	geen juiste th in het woord <i>thin</i>	lack of appropriate th in <i>thin</i>
THOMAS	geen juiste th in de naam <i>Thomas</i>	lack of appropriate th in <i>Thomas</i>
VAN	geen verschil tussen de eerste medeklinkers in <i>van</i> en <i>fan</i>	no distinction between the initial consonants in <i>van</i> and <i>fan</i>
WINE	geen verschil tussen de eerste medeklinkers in <i>wine</i> en <i>vine</i>	no distinction between the initial consonants in <i>wine</i> and <i>vine</i>

The errors illustrated by ANNIE, BAT and EXAM all revolve around the Dutch resistance to any realisations of English /æ/ that are both front and open – probably as the combined result of the absence of a similar sound in the vowel inventory of Standard Dutch (even though a highly stigmatised [æ] may be heard in for example the basilectal Utrecht urban accent) and antiquated pedagogical norms, on the basis of which “old-fashioned closer types of /æ/” are prescribed (Collins & Mees 2003b: 94). Consequently, *Annie* and *bat* are pronounced as what would be perceived by native speakers as /eni/ and /bet/. Dutch learners may also avoid [æ] in *exam*, pronouncing it as [ɪgʒɑ:m], which some assume to be the RP pronunciation (in a false analogy to the “BATH” words; see Wells 1982: 133; Collins & Mees 2003b: 120, 288). Since native speakers only pronounce *exam* with /æ/, the error of “rhyming *exam* with *jam*” will have the effect of a distractor. It should also be noted that since *Annie* and *penny*, having different initial consonants, are not minimal pairs, the error of /æ ~ e/ conflation may be less obvious to non-native judges of this token than in

the case of *bat* and *bet*. According to Collins & Mees (2003b: 94), the error is “notoriously persistent” in the English of Dutch learners.

Other phonemic errors involving vowel contrasts are COLOUR, PULL and SURE. The first two represent well-known pronunciation problems of Dutch learners, which are both described by Collins & Mees (2003b: 97, 290) as “persistent”. While COLOUR represents the neutralisation of the /ʌ ~ ɒ/ contrast as a result of a spelling pronunciation, PULL is an example of the conflation of /ʊ/ and /u:/, two vowels which, according to Collins & Mees (2003b: 97) are confused by “[a]ll Dutch-speaking students”. The error may well be particularly salient because the conflation is illustrated by a minimal pair (*pull* versus *pool*). Conversely, SURE is only likely to be an error, if at all, for those few speakers of what Wells (1982: 162–163) terms “conservative RP”. Such speakers may conceivably object to the modern RP realisation of “CURE” words with /ɜ:/, as a result of which *sure* has become homophonous with *shore* (see also Collins & Mees 2003b: 114). Although this will not be true of all varieties of English (including GA), in modern-day RP the conflation of *sure* and *shore* (already noted by Rippmann 1906: 69–70) is unlikely to be considered an error. From this point of view, SURE may be considered, if not a distractor, then at least a litmus test of how up-to-date the Dutch judges’ pedagogical norms are if modern RP is their norm.

The test contained seven phonemic errors involving consonant contrasts, three of which represented the neutralisation of the fortis/lenis contrast in final position (BED and OFF) and in word-initial position (VAN). According to Collins & Mees (2003b: 290), these are “persistent” errors in the L2 English of Dutch learners. The former, known as “final devoicing” (or by the German term *Auslautverhartung*) in the Dutch phonetic-phonological literature, is a process that, similarly to German, Polish, Russian and many other languages, results in the devoicing (“fortis” pronunciation) of final obstruents, cf. *bond* [bont] “union” (plural *bond-en*), *bont* [bont] “colourful” (inflected form *bont-e*), and *krap* [krap] “narrow” (inflected form *krapp-e*), *krab* [krap] “scratch” (infinitive *krabb-en*) (see Trommelen & Zonneveld 1979: 51, Booij 1977: 175). The latter is a phenomenon especially persistent in Western Dutch, by which the opposition between voiced and voiceless (lenis and fortis) initial fricatives is lost (see Bremmer & Gussenhoven 1983, Booij 1995: 7–8). Another such phonemic error is the “[c]onfusion of /v-w/ contrast” as illustrated by WINE: in (most Netherlands) Dutch, /w/ is realised as [v], which is perceived by English speakers as /v/ (Collins & Mees 2003b: 290, 174–176). The presentation of these errors as resulting in the confusion of minimal pairs (*bed* ~ *bet*, *of* ~ *off*, *van* ~ *fan*, *wine* ~ *vine*) will have made them particularly salient. In addition, there were three errors illustrating TH-stopping (replacement of dental fricatives by dental plosives or affricates). One of these was a distractor, as *Thomas* is not actually pronounced with /θ/. As Collins & Mees (2003b: 142) note, “[r]eplacement of /ð/ by /d/”, as illustrated by *THAT*, “is one of the most common and persistent Dutch errors”. A possibly less persistent, but equally significant error is the substitution of /θ/ by /t/ (Collins & Mees 2003b: 291), which has now

overtaken the former tendency to replace /θ/ by /s/ (still noted by Collins & Mees 2003b: 142). Dutch, in common with the overwhelming majority of the languages of the world, lacks the dental fricatives /θ, ð/.

Some **realisational** errors (see Table 2.2 for an overview) simply represent L2 pronunciations that may be associated with infrequent or stigmatised L1 realisations, such as the use of uvular-**r** in RED, which, according to Collins & Mees (2003b: 287, 179) is “completely unacceptable”, especially the “very strong, scrappy uvular fricative [ʁ] (sometimes termed “brouwende r”)” heard both in some Southern regional accents of Dutch and in “affected types of ABN”.⁷ This is “unpleasant to English ears” (Collins & Mees 2003b: 179). Similarly, an overdark, pharyngealised [ɤ], as in FULL, may “sound ugly to an English ear” (Collins *et al.* 1987: 30) or be associated with L-vocalisation (see 3.5.20 and 4.2.19). As Collins & Mees (2003b: 170–171) note, “Netherlands Dutch dark [ɤ]” differs significantly “from its English counterpart” in that the Dutch realisation involves “pharyngealisation rather than velarisation with a noticeable retraction of the tongue-root towards the pharynx wall”, often combined with the absence of any “contact between the tongue and the alveolar ridge, so that the articulation takes on the character of a back vowel”. Other realisations may be less noticeable, including the use of Dutch /h/ in English, as in hot. According to Collins & Mees (2003b: 148), the difference between English /h/, which is “only voiced between voiced sounds”, and Dutch /h/, which “is more likely to have voice in all contexts” is “not easily perceived by English native speakers”. This implies that HOT may serve as a useful distractor.

Other realisational errors may be considered particularly important since some of the realisations in question may create the effect of phoneme substitution, as in the case of DEAD, ICE and TIN. (These errors were all described as “most significant” or “significant” by Collins & Mees 2003b: 290–291.) For instance, the error in DEAD represents the replacement of a lenis consonant (/d/) by a glottal stop, as in *[deʔ] for *dead*, which will be perceived by native speakers as *debt*. This is because, in native English, such glottal substitutions almost exclusively involve fortis stops such as /t/ (Collins & Mees 2003b: 153, see also 4.2.10 and 4.4.10). Glottal substitution is not a feature of Dutch, and Dutch learners tend to overgeneralise this to include lenis stops (as with the “overgeneralisation of preglottalisation”, described by Gussenhoven & Broeders 1997: 131 as “an inevitable stage in the learning process”). Similarly, the significance of over-long /aɪ/ before fortis consonants (Collins & Mees 2003b: 290), as in ICE, lies in this word being misinterpreted as *eyes* (see 3.5.9). In Dutch, the /a/ vowel is long in this sequence compared to its English counterpart (cf. *m/a(:)ɪ/s* “maize”). Likewise, because of the similarities between word-initial Dutch unaspirated /p, t, k/ and word-initial English devoiced /b, d, g/, the absence of aspiration in TIN may result in the word being misinterpreted as *din* (Gussenhoven & Broeders 1997: 129). In the case of THAT_MAN, however, the

⁷ *ABN* (*Algemeen Beschaafd Nederlands*) is one of the terms commonly used to denote the prestige variety of Dutch in the Netherlands.

use of a glottal stop, either as reinforcement or as replacement of the final /t/ in *that*, is very unlikely to cause native speakers to hear a different phoneme. In this position, glottal reinforcement is in fact consistent with RP and GA norms (Collins & Mees 1993: 14, 2003b: 153). It may be noted that glottal substitution is also increasingly common, even in certain environments in RP (Wells 1982: 261). This suggests that *THAT_MAN*, rather than being a significant error, may well serve as a useful distractor.

Table 2.2. Descriptions in Dutch of the **realisational** errors as presented to participants in the Dutch Experiment, arranged in alphabetical order according to keywords, and with English translations.

Key word	Dutch description of error	English translation
DEAD	een glottal stop zeggen in <i>dead</i>	producing a glottal stop in <i>dead</i>
FULL	geen juiste donkere l gebruiken in <i>full</i>	lack of appropriate dark l in <i>full</i>
HOT	een Nederlandse h gebruiken in <i>hot</i>	using a Dutch h in <i>hot</i>
ICE	de klinker in <i>ice</i> even lang maken als die in <i>eyes</i>	no difference in vowel length between <i>ice</i> and <i>eyes</i>
RED	een huig- r gebruiken in <i>red</i>	using a uvular- r in <i>red</i>
TIN	geen aspiratie in <i>tin</i>	no aspiration in <i>tin</i>
THAT_MAN	een glottal stop zeggen in <i>that man</i>	producing a glottal stop in <i>that man</i>

An overview of the **distributional** errors in the Dutch Experiment is provided in Table 2.3. These include three errors illustrating problems with **r**-distribution (CAR, FARMER, IDEA, INDIA). Accents of English are divided into *rhotic* (or “r-ful”) accents, where /r/ is almost invariably pronounced as indicated in the orthography, and *non-rhotic* (or “r-less”) accents, in which /r/ is not pronounced in certain contexts. While such **r**-deletion or “R-dropping” is the norm in a non-rhotic accent such as RP, which only retains pre-vocalic /r/, this phoneme is also sounded pre-consonantly and pre-pausally in a rhotic accent such as GA (Wells 1982: 75–76, 218ff). This means that “pronouncing **r** in the word *car*” would be a significant error in RP (Collins & Mees 2003b: 291) but the expected norm in GA (Collins & Mees 1993: 33). However, the opposite would be true for “not pronouncing **r** in *farmer*”. While both **rs** are sounded in GA, neither would be pronounced in RP.

If particular pronunciations are potential errors in one major variety of Standard English (such as RP and GA) but are consistent with the norm in another, this will be referred to as *mirroring*, to serve as a reminder that respondents’ judgements could be influenced by their attitude to the presence or absence of such realisations in the other main variety. This may, for instance, be relevant in those cases where respondents only object to particular

pronunciations *because* they are associated with varieties such as RP or GA (as will be discussed below).

Table 2.3. Descriptions in Dutch of the **distributional** errors as presented to participants in the Dutch Experiment, arranged in alphabetical order according to keywords, and with English translations.

Key word	Dutch description of error	English translation
CAR	de r uitspreken in het woord <i>car</i>	pronouncing r in the word <i>car</i>
FARMER	geen r uitspreken in <i>farmer</i>	not pronouncing r in <i>farmer</i>
FILM	een klinker uitspreken tussen de l en m in <i>film</i>	pronouncing a vowel between l and m in <i>film</i>
HOT_TEA	slechts één t in <i>hot tea</i> uitspreken	pronouncing only one t in <i>hot tea</i>
IDEA	een r uitspreken voor het woord <i>of</i> in de frase <i>idea of it</i>	pronouncing r before the word <i>of</i> in the phrase <i>idea of it</i>
INDIA	<i>India</i> laten rijmen met <i>windier</i>	rhyiming <i>India</i> and <i>windier</i>
NEW	geen j in het woord <i>new</i>	no j in the word <i>new</i>
SUIT	geen j in het woord <i>suit</i>	no j in the word <i>suit</i>

Another potential error in the distributional group which may be subject to the effects of *mirroring* is INDIA. While *India* is a rhyme (or near-rhyme) with *windier* in RP, needless to say this is not the case in GA. Similarly, although intrusive-**r** in *idea of it* is, while stigmatised, “a regular feature of RP” (Collins & Mees 2003b: 178), this feature is not found in GA (but “common” in “New York speech” and some other non-rhotic accents of North America, cf. Wells 1982: 507, 520, Hay & Sudbury 2005: 801). This implies that intrusive-**r** in IDEA would be an unmistakable error by GA standards, while only some judges would be likely to see it as such within the context of RP. The latter include those native speakers of British English whose “strong reaction” to intrusive-**r** prompted Collins & Mees (2003b: 179, 181) to advise non-native learners against using it.

If respondents base their answers on what would be expected for RP, they are likely to treat FARMER, INDIA and possibly IDEA as distractors, whereas if GA is their model, this would only be true of CAR. It was assumed that, since RP is the most commonly taught variety of English in the Netherlands (Van der Haagen 1998: 2), respondents’ answers would tend to be consistent what would be expected if RP was the model. To some extent, this is actually borne out by the fact that all groups of respondents assigned much more importance to CAR than to INDIA, FARMER or IDEA (see 2.3). Nevertheless, in retrospect it would have been useful to verify this by asking judges explicitly which model they had referred to when assessing the severity of the 40 errors in question.

There are two other tokens that can only be classified as distributional errors if they are judged against a particular pronunciation model. These are *NEW* and *SUIT*, neither of which are mentioned in Collins & Mees (2003b), or any other pronunciation textbook aimed at Dutch learners, as being significant errors. The potential errors in *NEW* and *SUIT* both revolve around the uncertainty Dutch learners may have about the use of a non-orthographic palatal glide or “yod”, which, in certain contexts, may or may not be required in different varieties of English. For instance, the failure to pronounce a palatal glide after /n/ in *new* would only be an error from the point of view of RP, but clearly not in GA, as in this context yod is not actually pronounced by the vast majority of speakers (Wells 1982: 247). Its inclusion in the experiment is warranted by the symbolic value it may have for different groups of native and non-native speakers as a tokenist representation of British versus American pronunciation (see 3.5.16). Apart from the fact that, in initial clusters, there is no corresponding use of non-orthographic yod in Dutch, it should be noted that Dutch completely lacks initial /nj-/ sequences. Although English /sj-/ sequences are unproblematical for most Dutch learners (Collins & Mees 2003b: 147), Dutch learners may be confused by the “variability” of yod-dropping in accents such as RP, where *suit* may be pronounced either as /sju:t/ or as /su:t/ (Wells 1982: 207).⁸ While it would be theoretically possible for very conservative speakers of RP to object to the latter, yod-less, realisation of *suit* as a serious error, this would be unlikely in view of the fact that the overwhelming majority of British people, including most RP speakers, pronounce it in this way (Collins & Mees 2003b: 146, Gussenhoven & Broeders 1997: 156, Wells 1982: 207, 2000: 748). If the latter pronunciation is not an error in modern RP, this implies that the intended error in *SUIT* is likely to serve merely as a distractor, to be assessed severely only by those respondents who adhere to what would appear to be outdated descriptions of RP.

Other distributional errors include schwa epenthesis in *FILM* and degemination in *HOT_TEA*. The former is described by Collins & Mees (2003b: 171) as an “unacceptable” Dutch error which is nevertheless found “in a few English dialects (e.g. types of Scottish, Irish, Lancashire)”. Schwa epenthesis in such final clusters is a well-known feature of Standard Dutch, as in *fī* [I^om] “film”, *he* [I^op] “help” (see Trommelen 1983: 77, Collins & Mees 2003b: 171). Degemination is a Dutch L2 error which is described by Collins & Mees (2003b: 218) as the reduction of “sequences of identical consonants, by elision, to a single consonant”, which is “a significant problem when it is imposed on sequences of plosives”. In Dutch, degemination is a general process applying to sequences of identical consonants in any position (see Trommelen & Zonneveld 1979: 108).

⁸ Additional confusion may arise as a result of the fact that Dutch does not maintain a difference between initial /sj-/ and /ʃ-/ , as a result of which some Dutch learners of English may pronounce *suit* incorrectly as “*/ʃu:t/” (Collins & Mees 2003b: 146).

The four **stress** errors included in the Dutch Experiment are presented in Table 2.4. These are intended as examples of what Collins & Mees (2003b: 291) refer to as the “misplacement of primary stress”, which they describe as one of the “most significant” (but “non-persistent”) errors of Dutch learners. Two of the four potential errors included here may be classified as stress errors in all varieties of English. IMAGIN exemplifies the failure to retain stem stress before the suffix sequence *-ative*, possibly as a result of learners’ over-application of English stress-shift as triggered by adjectival suffixes (e.g. ‘*adjective* ~ *adjectival*), or as an analogy to *imagination*. Similarly, incorrect stress placement in PERFECT illustrates the learner’s failure to employ what Collins & Mees (2003b: 231) term a “switch stress pattern” in English. Instead of stress being placed on the prefix, as is required for the adjective, learners place the stress incorrectly on the second syllable, which would be the correct stress for the Dutch adjective *per'fect* “perfect”, but which in English is associated with the corresponding verb. Both PERFECT and IMAGIN exemplify mistakes commonly made by Dutch learners. While DEVIANT is a transparent case of a distractor based on spelling (since, clearly, *deviant* and *defiant* do not actually rhyme at all), ADVERT is unlikely to be considered an error by native speakers unless their model is something other than North American English and they reject alternative British but “non-RP” pronunciations such as /ædvə'taɪz mənt/ (see Wells 2000: 12). Thus, reservations about the potential error in ADVERT may be construed as objections to non-standard or trans-Atlantic forms.

Table 2.4. Descriptions in Dutch of the **stress** errors as presented to participants in the Dutch Experiment, arranged in alphabetical order according to keywords, and with English translations.

Key word	Dutch description of error	English translation
ADVERT	klemtoon op de lettergreep <i>tise</i> in <i>advertisement</i>	stressing the syllable <i>tise</i> in <i>advertisement</i>
DEVIANT	<i>deviant</i> en <i>defiant</i> niet laten rijmen	failure to rhyme <i>deviant</i> and <i>defiant</i>
IMAGIN	klemtoon op de lettergreep <i>nat</i> in het bijv. nw. <i>imaginative</i>	stressing the syllable <i>nat</i> in the adj. <i>imaginative</i>
PERFECT	klemtoon op de lettergreep <i>fect</i> in het bijv. nw. <i>perfect</i>	stressing the syllable <i>fect</i> in the adj. <i>perfect</i>

Finally, Table 2.5 lists eight potential errors connected with **supra-segmental** phenomena such as intonation, contraction and weakening. It was difficult to describe these errors to the respondents without using at least some specialist terms such as “contracted forms” and “weak forms”. The latter term was actually provided in English because it was assumed that respondents would be more familiar with this than any Dutch equivalent. Another supposition was that the detailed technical descriptions required to incorporate specific examples

of typical L2 intonation patterns would place an unacceptably high interpretative burden on the respondents. This is why the only intonation token included in the Dutch Experiment was the very general problem of “too little variation in intonation”. According to Collins & Mees (2003b: 291), this is a “significant error”. The same authors also emphasise the significance of weak and contracted forms as potential sources of error. As contracted forms are “essential in spoken English” and should be used frequently, the “error” of frequent use of contracted forms (FREQ_CFS) is clearly a distractor, while their infrequent use (INF_CFS) is among the “most significant” errors made by Dutch learners (Collins & Mees 2003b: 20, 290). Almost precisely the same would be true of the distractor FREQ_WFS (the frequent use of weak forms) as opposed to the very salient error of INF_CFS (the infrequent use of weak forms), except that, as Collins & Mees (2003b: 20) suggest, the avoidance of contracted forms “is perhaps even more immediately noticeable” than that of weak forms. Nevertheless, while the latter “also play an important part in Dutch”, as Collins & Mees (2003b: 20) note, their infrequent use in L2 English “is one of the main sources of error for Dutch-speaking students”. This is why it was decided to include two further examples of the avoidance of weak forms (FROM and TO_WALES) in the Dutch Experiment.

Table 2.5. Descriptions in Dutch of the **suprasegmental** errors as presented to participants in the Dutch Experiment, arranged in alphabetical order according to keywords, and with English translations.

Key word	Dutch description of error	English translation
FREQ_CFS	veelvuldig gebruik van samen-trekkingen zoals <i>can 't, you 'll, I've, enz</i>	frequent use of contracted forms such as <i>can 't, you 'll, I've, etc.</i>
FREQ_WFS	veelvuldig gebruik maken van zg <i>weak forms</i>	frequent use of so-called <i>weak forms</i>
FROM	<i>from</i> laten rijmen op <i>Tom</i> in de zin <i>where does he come from?</i>	rhyiming <i>from</i> and <i>Tom</i> in the phrase <i>where does he come from?</i>
INS_CFS	onvoldoende gebruik van samen-trekkingen zoals <i>can 't, you 'll, I've, enz</i>	insufficient use of contracted forms such as <i>can 't, you 'll, I've, etc.</i>
INS_WFS	onvoldoende gebruik maken van zg <i>weak forms</i>	insufficient use of so-called <i>weak forms</i>
INT	weinig gevarieerde intonatie gebruiken	too little variation in intonation
SECONDAR	<i>secretary</i> en <i>secondary</i> als vier lettergrepen uitspreken	pronouncing <i>secretary</i> and <i>secondary</i> with four syllables
TO_WALES	klemtoon op <i>to</i> in <i>going to Wales</i>	stressing <i>to</i> in <i>going to Wales</i>

In the contexts provided (i.e. *where does he come from?* and *going to Wales* respectively), the prepositions *to* and *from* are unlikely to attract the nuclear stress associated with the strong forms. As a result, one would expect to hear these weakened in L1 English, but not necessarily in the English produced by Dutch learners. However, it should be remembered that native speakers of English are more inclined to stress prepositions contrastively than would be true of Dutch learners (Collins & Mees 2003b: 280; see also 3.5.17). Consequently, the correct use of the strong form would be a remote possibility in TO_WALES, but not in FROM, since, according to Collins & Mees (2003b: 20) strong forms are always “used at the end of the intonation group”, regardless of whether the word in question is stressed or unstressed. This would make FROM a distractor, while TO_WALES is more likely to be considered a significant error.

The error exemplified by SECONDAR may also be described as avoidance of weakening, except that this case of lexical-incidential weakening, while independent of nuclear stress, is required in one major variety of English such as RP while absent in for instance GA. Although this potential error is not presented as important in any of the relevant textbooks, it may well be subject to the effects of mirroring and thus accorded more significance by those who believe, for instance, that a characteristically American realisation of *secondary* with four syllables is undesirable in the RP-modelled accents of Dutch learners. This in itself should warrant its inclusion.

As the above discussion will have demonstrated, potential errors may serve as distractors in a number of different ways. For instance, the errors described in DEVIANT, EXAM, FROM, INF_CFS, INF_WFS, THAT_MAN and THOMAS are in fact the negated versions of what all native speakers are likely to consider to be unacceptable. In other words, the **real** error would be to actually pronounce an “appropriate **th**” in *Thomas* or to rhyme *deviant* and *defiant*. It is to be assumed that reasonably competent judges of Dutch pronunciation errors would consider such distractors to be totally insignificant. Conversely, ADVERT, CAR, FARMER, FILM, IDEA, INDIA, NEW, SECONDAR, SUIT, SURE refer to realisations that would only be considered errors by particular groups of respondents, such as those objecting to characteristically American or British pronunciations or to what they perceive to be non-standard or stigmatised features. Since these realisations are considered problematical by at least some groups of native speakers, one would expect competent judges of Dutch pronunciation errors to rank such potential distractors more severely than the “negated versions” mentioned above. Naturally, this excludes those cases where the intended error is unlikely to be considered at all important by any group of native speakers (such as HOT, SUIT, SURE, and possibly ADVERT). However, if any Dutch respondents evaluated these as serious errors, this could be taken to mean that such judges adhere to obsolescent pronunciation norms (at least as regards SUIT and SURE). A short description of Dutch respondents’ evaluations of all these “potential distractors” is provided in 2.3.

2.1.4 Priorities in data analysis

The five versions of the Dutch survey discussed in 2.1.2 generated a wealth of data, and an analysis of these requires establishing priorities. In the context of this dissertation, the data obtained from the Dutch Experiment will be analysed for three purposes. These are: (1) to present the most striking results of a brief analysis of respondents' views on teaching English pronunciation, at a variety of levels, in educational settings in the Netherlands (to be provided in 2.2.1); (2) to help select the most significant errors for the core experiment involving native speakers (to be discussed in 2.3); and (3) to compare the assessments, by native and non-native respondents, of those potential errors that are similar or identical in both experiments (to be discussed in section 3.6).

2.2 Dutch respondents' views on teaching English pronunciation

2.2.1 Data analysis and discussion of the results

The groups of university and college lecturers, and the group of secondary school teachers, were all asked about the role of pronunciation in their own teaching or teaching environment; the groups of students and pupils were asked to place this in the context of their previous experiences in secondary school. This implies that the answers of the lecturers relate to the situation at universities and colleges, whereas those of the schoolteachers refer to the role of pronunciation in secondary schools – as do the answers of the students and pupils. It must be pointed out, however, that students and pupils will not necessarily share the perceptions of those actually engaged in teaching. This makes it necessary to divide the respondents into three main groups:

- (1) the 62 university and college lecturers (NL/LEC);
- (2) the 101 students and pupils (NL/STU);
- (3) the 98 secondary school teachers (NL/SST).

The data from these three groups cannot be fed into a single analysis, because they reflect different properties among these three groups. In consequence of this, only pairwise comparisons among groups (using χ^2 statistics) will be presented below, except where indicated otherwise. The relevant pairwise comparisons are (1) those between the two groups potentially engaged in teaching pronunciation, i.e. NL/SST and NL/LEC, and (2) those between the two groups describing actual teaching practice in secondary schools, i.e. NL/SST and NL/STU.

Apart from questions about pronunciation, respondents were also asked whether English was the medium used in classes. Table 2.6 shows that only

24% of the NL/SST group stated that they always taught in English, whereas this was true of no fewer than 94% in the NL/LEC group. The answers “occasionally” or “never” were found in 22% of the NL/SST respondents, but not once with those in the NL/LEC group. The differences in distribution between these two groups are significant ($\chi^2 = 74.134$, $df = 2$, $p < .001$). It should be remembered, however, that six out of 90 NL/SST reported that they were native or bilingual speakers of English, as opposed to 21 out of 62 NL/LEC. This is likely to have positively affected the lecturers’ willingness to use English in the classroom.

Table 2.6. Frequencies and percentages of respondents in the NL/SST and NL/LEC groups who stated that English was either “always”, “regularly” or “occasionally/never” used as a medium in the English classes they referred to.

	Always	Regularly	Occasionally / never
NL/SST	23 (24%)	53 (55%)	21 (22%)
NL/LEC	58 (94%)	4 (7%)	0 (0%)

As Table 2.7 shows, as many as 37 percent of the NL/STU group reported that the secondary school teachers who taught them English had hardly ever used this language as a medium of instruction. The differences between the NL/STU and NL/SST groups were significant ($\chi^2 = 7.872$, $df = 2$, $p = .02$). It is not improbable that the students (and pupils) are more critical as a group than the schoolteachers. After all, some students may actually have been motivated to take part in the experiment as a means of criticising the way in which they were taught English by their former teachers, whereas this is unlikely to have been a factor in the case of the secondary school teachers, who had volunteered to report on their own teaching. Nonetheless, a clear trend emerges whereby English is used considerably less often in Dutch secondary schools than in universities and colleges. Within the context of pronunciation, this implies that secondary school pupils taking English classes are exposed much less to spoken English (whether native or non-native) than students reading English at universities and colleges.

Table 2.7. Frequencies and percentages of respondents in the NL/SST and NL/STU who stated that English was either “always”, “regularly” or “occasionally/never” used as a medium in the English classes they referred to.

	Always	Regularly	Occasionally / never
NL/SST	23 (24%)	53 (55%)	21 (22%)
NL/STU	11 (12%)	49 (52%)	35 (37%)

Not only do the secondary school teachers in this survey tend to use English less frequently as a teaching medium than the lecturers, but they also believe that they focus on pronunciation more frequently than do the latter. Only one out of 97 respondents in the NL/SST group stated they paid “no, or hardly any” attention to pronunciation, as opposed to 20 out of 62 in the NL/LEC group (i.e. 32%). It should of course be noted that all secondary school teachers are likely to be involved in teaching proficiency, unlike some of the college and university lecturers (see also 3.6). In addition, the interest in pronunciation teaching professed by the teachers who volunteered to take part in the survey may not be shared by those who did not choose to participate.

In this respect, the contrast between the schoolteachers on the one hand, and the students and pupils on the other, is particularly noticeable. No fewer than 37 out of 95 Dutch students (i.e. 39%) indicated that they had received little or no pronunciation training in secondary school. Moreover, as Table 2.8 shows, 74% of those NL/STU participants who specified the frequency of this training referred to it as “occasionally”, as opposed to 37% of the relevant NL/SST respondents who stated this. The differences in distribution between these two groups are significant ($\chi^2 = 34.832$, $df = 2$, $p < .001$).

Table 2.8. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who specified that pronunciation training was either given “in every class”, “every week” or “occasionally” in the English classes they referred to.

	In every class	Every week	Occasionally
NL/SST	44 (46%)	16 (17%)	35 (37%)
NL/STU	5 (6%)	15 (19%)	57 (74%)

Inasmuch as these students’ observations can be taken to be representative of the general situation in the Netherlands, this would seem to indicate that very little direct attention is being paid to English pronunciation training in secondary schools. In addition, even the teachers who took part in this survey typically did not think that pronunciation was an especially important subject. A three-way comparison shows that in this respect, their views differed very significantly from those generally held by the NL/STU and NL/LEC groups.

As can be seen in Table 2.9, only 29% of the NL/SST participants considered pronunciation training to be either “essential” or “very important”, whereas 56% viewed it as “a normal part of the course” (Dutch “*is gewoon één van de onderdelen*”). The percentages of participants who described pronunciation training as “essential” or “very important” were in fact more than twice those for the NL/LEC and NL/STU groups. These differences were statistically significant ($\chi^2 = 30.667$, $df = 4$, $p < .001$). If the attitudes of these secondary school teachers (virtually all of whom state they teach pronunciation at least

occasionally) are already so very distinct from the other two groups, one might well wonder how much importance would be attached to such training by those Dutch teachers who did not participate in this survey. Those who did take part may have been more than usually motivated to emphasise the importance of pronunciation training and, as a result, may have over-reported an interest in pronunciation which is in fact less marked than that of the NL/LEC and NL/STU groups. In addition, it would also be interesting to discover how English pronunciation training would be rated by university students not reading English language and literature. That is to say, the students of English in the NL/STU group may have been more inclined to acknowledge the importance of English pronunciation training than other students. This would be an interesting avenue for further investigation.

Table 2.9. Frequencies and percentages of respondents in the NL/SST, NL/LEC, NL/STU groups who stated that pronunciation training was either “essential” or “very important”, “a normal part of the course” or either had “limited use” or “no use at all”.

	Essential / very important	A normal part of the course	Has limited use, or no use at all
NL/SST	28 (29%)	54 (56%)	15 (15%)
NL/LEC	42 (68%)	16 (26%)	4 (6%)
NL/STU	59 (62%)	28 (29%)	8 (8%)

Arguably, the low priority given to pronunciation training in Dutch secondary schools can be further illustrated by the limited importance attached to pronunciation models. It may be assumed that any kind of detailed attention to pronunciation training would necessarily involve the conscious adoption of a particular model, even if it is some kind of International English as proposed by Jenkins (2000). However, no fewer than 47% of the teachers stated that their school did not prescribe or use any particular pronunciation models, whether RP, GA, or any other. Yet only 16% of the lecturers stated that no model was used in the departments or colleges in question (see Table 2.10). These groups' distributions are significantly different ($\chi^2 = 16.376$, $df = 2$, $p < .001$).

The limited significance attached to such models could either imply a lack of interest in pronunciation on the part of the schoolteachers, or, as is evident from a small number of comments, reservations about the feasibility of teaching particular groups of pupils a native variety of English. Two teachers even questioned the idea of pronunciation models in view of what they felt was the emergence of a “global” or “European” variety of English. However, this was not evident from any of the other responses from the NL/SST group. It is also striking that 43% of the teachers stated that their schools prescribed RP.

This tendency to prescribe “either RP or nothing” is much less clearly evident from the lecturers’ observations about the English model employed in English departments of Dutch universities.

Table 2.10. Frequencies and percentages of respondents in the NL/SST and NL/LEC groups who stated that the pronunciation model used in their schools or departments was either “none”, “RP”, or “other than RP” (i.e. either GA, both RP and GA, or a different accent).

	None	RP	Other than RP
NL/STT	46 (47%)	42 (43%)	9 (9%)
NL/LEC	10 (16%)	38 (62%)	13 (21%)

A pairwise comparison of respondents in the NL/SST and NL/STU groups reveals that students and pupils did not respond to the question of pronunciation norms significantly differently ($\chi^2 = 4.198$, $df = 2$, $p = .123$). As Table 2.11 shows, only 38% of the latter indicated that a pronunciation norm (“RP” or “other than RP”) was implemented. It should be noted that some students may not have been aware that a particular pronunciation model was being employed, since this had not been made explicit by their teachers. A tendency for teachers to avoid explicit reference to any pronunciation norms is also evident from the way participants in the NL/SST group described the frequency with which teachers referred to a particular pronunciation model.

Table 2.11. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who stated that the pronunciation model used in secondary schools was either “none”, “RP”, or “other than RP” (i.e. either GA, both RP and GA, or a different accent).

	None	RP	Other than RP
NL/STT	46 (47%)	42 (43%)	9 (9%)
NL/STU	59 (62%)	30 (32%)	6 (6%)

Table 2.12 shows that while the vast majority of teachers, students or pupils felt that infrequent reference was made to pronunciation models, only one out of 94 students felt that this was done on a regular basis, as opposed to 14 out of 93 schoolteachers (i.e. 15%) who stated this to be so. The differences in distribution are indeed significant ($\chi^2 = 13.829$, $df = 2$, $p < .001$).

Table 2.12. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who stated that the pronunciation model used was referred to “regularly”, “sometimes” or “hardly or not at all” during the English classes in question.

	Regularly	Sometimes	Hardly or not at all
NL/SST	14 (15%)	36 (39%)	43 (46%)
NL/STU	1 (1%)	34 (36%)	59 (63%)

A similar pairwise comparison of the NL/SST and NL/LEC groups (see Table 2.13) shows that references to any pronunciation model were reported significantly less frequently for secondary schools than for universities ($\chi^2 = 14.843$, $df = 2$, $p < .001$). To summarise, pronunciation models were neither employed nor explicitly referred to as frequently in schools as they were in universities. If any model was used in secondary schools, this was RP rather than any other.

Table 2.13. Frequencies and percentages of respondents in the NL/SST and NL/LEC groups who stated that the pronunciation model used was referred to either “regularly”, “sometimes”, or “hardly or not at all” during the English classes in question.

	Regularly	Sometimes	Hardly or not at all
NL/SST	14 (15%)	36 (39%)	43 (46%)
NL/LEC	26 (43%)	14 (23%)	21 (34%)

Participants in all five versions were also asked to indicate to what extent certain activities and materials were used to help pupils and students improve their pronunciation. While this yielded no striking differences between the NL/SST and NL/LEC groups, there were three particularly salient instances where the responses of NL/SST and NL/STU groups were significantly different. For instance, while 74% of the schoolteachers stated that they had actually undertaken an activity such explaining the differences between Dutch and English pronunciation, 69% of the students and pupils intimated that their teachers had not actually done this at all. The differences between the two groups are significant in this respect ($\chi^2 = 36.763$, $df = 1$, $p < .001$). These findings as presented in Table 2.14 imply that contrastive analysis of Dutch and English pronunciation features less often in English classes than would be expected on the basis of the classroom activities reported by the teachers. It is, however, theoretically possible that students may have under-reported such activities.

Table 2.14. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who either denied or confirmed that explaining the differences between Dutch and English pronunciation was one of the activities undertaken by English teachers.

	No	Yes
NL/STT	25 (26%)	72 (74%)
NL/STU	66 (69%)	29 (31%)

Another unexpected result is that no fewer than 19% of students and pupils reported that teachers did not normally make their students speak English as a classroom exercise. Conversely, only four percent of the teachers did not include this among the activities listed (see Table 2.15). The different distributions are significant ($\chi^2 = 10.395$, $df = 1$, $p < .001$), but, as in the other cases discussed, it is unclear whether it is the students who are under-reporting any exercise or training connected with pronunciation or the teachers are over-reporting these. Nevertheless, it may come as a surprise to learn that this activity, by certain of the teachers' own admission, is not routinely undertaken by all teachers at all levels. While some of the schoolteachers' comments suggest that this kind of training may be avoided so as not to alienate or discourage any pupils, especially those involved in pre-vocational education, it could also be argued that exempting pupils from a core activity of language learning actually means doing them a disservice. Moreover, some pupils may even enjoy speaking English.

Table 2.15. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who either denied or confirmed that making pupils speak English was one of the activities undertaken by English teachers.

	No	Yes
NL/STT	4 (4%)	93 (96%)
NL/STU	18 (19%)	77 (81%)

If, as the NL/STU responses appear to suggest, there is in fact a minority of Dutch teachers of English who do not actually make their pupils speak English as a classroom exercise at all, this should be a serious cause for concern. A point for future investigation might be the extent to which this is the result of overburdened programmes and unrealistic class sizes, as was pointed out by some respondents. In any case, it makes one wonder where the 87% of Dutch citizens who claim to "speak English" (European Commission 2005: 4) learn to do this. In this context, it would be tempting to speculate how many

Dutch people believe that ample exposure to subtitled television programmes in English has equipped them adequately for interaction in this language.⁹

In view of the emerging evidence that pronunciation is accorded so little priority in secondary schools, it is perhaps remarkable to learn that pupils' pronunciation is in fact assessed in 83% of the cases reported by the teachers. It would be expected that any subject that is widely subject to assessment is firmly anchored in actual teaching programmes. As Table 2.16 shows, as many as 67% of the students and pupils even reported that their pronunciation was graded in the secondary schools they had attended – although the distributions of the two groups are significantly different ($\chi^2 = 7.023$, $df = 1$, $p < .001$). It may be noted that 42 out of 96 teachers (44%), but only 20 out of 93 students and pupils (22%), stated that pronunciation was assessed in all years – although this may have been subject to over- or under-reporting. Some students also observed that their pronunciation had been judged even though the subject had not been taught. If almost half of the teachers report that pupils' pronunciation is subject to evaluation in all years – if only, as is pointed out in some of the comments, as part of the assessment of proficiency – this would suggest that this subject should be taught more widely than appears now to be the case.

Table 2.16. Frequencies and percentages of respondents in the NL/SST and NL/STU groups who either denied or confirmed that secondary school pupils' English pronunciation was assessed at some stage in the curriculum.

	No	Yes
NL/STT	16 (17%)	80 (83%)
NL/STU	31 (33%)	62 (67%)

Morley (1996: 146) lists a number of what she refers to as “myths of *misguided conventional wisdom*” which are used “as reasons for denying students access to the [speech and pronunciation] training they need”. These include “pronunciation isn't important”, “students will pick it up on their own”, “pronunciation is too hard to teach” and “I don't have the training to teach it, so I just won't bother (And I'll just say pronunciation isn't important)” (Morley 1996: 146–147). While the first three of these claims appear to be in evidence from both participants' responses and their comments, the last is understandably less easy to ascertain. (In addition, Morley's remark could almost be considered

⁹ In the Netherlands, access to cable television is among the highest in the world (Information Society Promotion Office of the European Commission 1999), and television channels such as BBC1, BBC2 and MTV are immediately accessible to almost all viewers. As in the Scandinavian countries and Belgium, virtually all channels (both domestic and foreign) aimed at viewers in the Netherlands provide subtitling of foreign-language items rather than dubbing.

as an *ad hominem* attack on those who seriously believe that pronunciation is unimportant.) Still, it raises the question to what extent teacher training may have prepared them inadequately for the task of pronunciation training. The present survey did generate at least one potentially useful indication of a link between pronunciation teaching and teachers’ personal background in this subject, namely the matter of teachers’ own pronunciation of English.

Of the three main groups, only the NL/STU described the accents of those who had taught them English in secondary school. They were asked to respond to the question: “Do you feel that your English teachers mostly have an easily recognizable British, American or other English accent?” (Dutch “*Heb je de indruk dat je leraren Engels over het algemeen met een duidelijk herkenbaar Brits, Amerikaans of ander Engels accent spreken?*”). The objection could possibly be raised that some Dutch students of English cannot distinguish the difference between a British accent and an American one – even though they have full access to the media. It would, however, be more difficult to maintain that they cannot detect a Dutch accent in English. This implies that some significance should be attached to the results (as presented in Table 2.17), which show that no fewer than 14% of students and pupils stated that their teachers’ accents were “mostly Dutch”.

Table 2.17. Frequencies and percentage of students and pupils’ descriptions of their secondary school teachers’ accents in English.

Mostly British	Mostly American	Some British; some American	Mostly another accent of English	Mostly Dutch	Didn’t notice
64 (68%)	1 (1%)	7 (7%)	1 (1%)	13 (14%)	8 (9%)

Arguably, all professional non-native teachers of a language should have a convincing command of that language in all its aspects, or at least be able to convey this impression to their pupils. If this requirement includes avoidance of clearly noticeable L1 interference, it could be argued that 14 percent is indeed high. To the extent that these results are a representative sample of secondary school teachers of English in the Netherlands, it would suggest that, unless one advocates a deliberate policy of teaching non-native English, at least some teachers could do with remedial pronunciation training themselves. This is one of the suggestions made in 2.2.2, where a number of the preliminary conclusions and recommendations are provided, based on the results discussed in this section.

2.2.2 Preliminary conclusions and recommendations

In conclusion, a brief analysis of respondents’ views on teaching English pronunciation shows that instructors giving English classes in Dutch secondary

schools employ the medium of English considerably less often than do their counterparts giving English lectures and tutorials in universities and colleges. In addition, secondary school teachers require their students to speak English less frequently than might reasonably be expected – in some cases not at all. Furthermore, pronunciation is taught less often and considered much less important in secondary schools than at universities, even though most teachers indicate that students' pronunciation is in fact assessed at some stage in the curriculum. Moreover, secondary schools employ pronunciation models less often, and teachers discuss them less frequently, than do English departments of universities and colleges. While most teachers indicated that they do discuss differences between Dutch and English pronunciation, nevertheless a great many students observed that this was not actually true of the secondary school teachers who had taught them.

If one accepts that a subject that is to be eventually evaluated should be taught, this should be reason in itself to recommend that secondary school teachers should pay more attention to pronunciation training and to using English in the classroom. As will also be discussed in 3.6, teachers would also do well to make more explicit reference to pronunciation models such as RP and GA. It should also be investigated to what extent circumstances beyond teachers' control (such as overburdened programmes and class size) contribute to the lack of focus on speech and pronunciation training. Another factor susceptible to investigation is the priority given to pronunciation in teacher training colleges – something that is especially relevant given that a number of students and pupils described their teachers' accents as “mostly Dutch” (Dutch *“Ze hadden meestal eerder een Nederlands accent”*.) This may be undesirable if native English continues to be regarded as an appropriate model for Dutch learners to imitate.

Those responsible for the English curricula in universities and colleges, including university administrators, need to be aware of the lack of focus on pronunciation and speech training as has been attested in a significant number of secondary schools. If they are concerned that university and college graduates should have reasonably convincing accents, they must ensure that pronunciation training continues to be firmly anchored in their programmes, or is even expanded – unless they assume that “students will pick it up on their own” (cf. Morley 1996: 146). A recommendation which follows is that they would do well to seek to establish cooperation between schools and universities in order to encourage pronunciation training at the level of secondary education – especially if they do not wish to spend valuable time in the first years of an English course helping students unlearn Dutch pronunciation habits unconsciously adopted at secondary school.

A general limitation of this part of the Dutch survey is that the different groups may have dissimilar motivations for taking part in this experiment. For instance, secondary school teachers may have been motivated to take part because of what may be an uncharacteristically positive attitude to pronunciation. On the other hand, students of English may have participated as

a means of criticising their former English teachers, or the extent to which the English curriculum in secondary schools may have failed to prepare them for an English course at university. As a result of this, teachers may have over-reported any activities associated with pronunciation, whereas students may have under-reported these. However, inasmuch as these groups can be represented as occupying such extreme positions, it may be prudent to assume that both groups' observations are equally valuable in establishing attitudes to pronunciation in Dutch secondary schools.

2.3 Pronunciation errors in the Dutch Experiment: data analysis

As was pointed out in 2.1.3, the second section of the Dutch survey consisted of a description of 40 possible pronunciation errors, which respondents were asked to rate on a five-point Likert scale, ranging from 0 (= no error) to 4 (= a very serious error). It was assumed that it is important to discover, within each of the five error categories, which errors were assessed as being among the most severe. It would then be possible to use this selection as the basis for a similar section on error assessment in the core experiment involving native speakers, which will be discussed in subsequent chapters.

If, for instance, Dutch respondents consider a characteristically Dutch error such as /æ ~ e/ conflation in BAT to be particularly serious, it would be useful to include this token in the core experiment in order to discover to what extent their assessment is shared by native speakers of English. Importantly, the selection of the most serious errors should be made for all five categories, so as to avoid giving undue weight to respondents' possible bias towards particular types of error (such as "phonemic" or "suprasegmental"). Clearly, an experiment in which certain error categories are over-represented could not be regarded as representative.

However, if participants in the Netherlands consider particular errors (such as "potential distractors", see 2.1.3) to be relatively insignificant, these are unlikely to provide an interesting basis for comparison with native-speaker judgements, unless there are indications in textbooks such as Collins & Mees (2003b) that respondents in the core experiment will view these very differently. Only in such cases will errors rated by Dutch judges as the least important in a particular category be incorporated into the Native-speaker Experiment.

The most significant errors in each of the five error categories may be selected on the basis of a type of weighted average to be referred to as the Error Severity Index (henceforth "ESI"). For each token, the ESI is obtained by adding up all severity judgements in a group of respondents according to the following formula: (number of assessments rated "no error" *1) + (number of assessments rated "a relatively unimportant error" *2) + (number of assessments rated "not very serious" *3) + (number of assessments rated "serious" *4)

+ (number of assessments rated “very serious” *5). These ratings correspond to Dutch “*geen fout*”, “*een vrij onbelangrijke fout*”, “*een minder ernstige fout*”, “*een ernstige fout*” and “*een zeer ernstige fout*” respectively. This weighted sum of assessments per token is subsequently divided by the total number of assessments in that group of judges, and multiplied by two to arrive at a figure between one and ten. The results for all five groups of Dutch respondents are presented in Table 2.18. The groups concerned are: the 98 secondary school teachers (NL/SST), the five secondary school pupils (NL/PUP), the 96 university students of English (NL/USS), the 52 university lecturers in English (NL/USL) and the ten college lecturers (NL/HBO).

Table 2.18. Error Severity Indices for all five versions by token (in alphabetical order) and by group of respondents.

	ADVERT	ANNIE	BAT	BED	CAR
NL/SST	7.39	5.81	7.67	8.31	5.79
NL/PUP	8.80	4.40	6.8	8.00	6.80
NL/USS	7.45	7.10	9.00	9.15	6.15
NL/USL	6.23	8.23	9.27	9.35	4.78
NL/HBO	6.40	7.80	8.80	9.00	5.60

	COLOUR	DEAD	DEVIANT	EXAM	FARMER
NL/SST	6.04	5.87	4.33	3.09	2.73
NL/PUP	8.40	2.00	3.00	4.00	3.20
NL/USS	6.28	7.95	3.49	3.15	2.67
NL/USL	7.77	6.80	3.36	2.27	2.24
NL/HBO	7.60	5.80	6.00	2.20	2.20

	FILM	FREQ_CFS	FREQ_WFS	FROM	FULL
NL/SST	8.02	2.96	3.68	4.14	4.63
NL/PUP	8.00	5.20	4.00	4.40	4.00
NL/USS	8.46	3.42	4.23	4.92	6.00
NL/USL	7.92	2.27	2.63	3.92	5.52
NL/HBO	6.89	2.40	2.60	3.00	5.20

	HOT	HOT_TEA	ICE	IDEA	IMAGIN
NL/SST	4.96	4.29	6.75	3.17	8.58
NL/PUP	5.20	5.20	6.40	4.50	7.20
NL/USS	5.29	3.77	7.98	3.05	8.53
NL/USL	4.00	4.27	7.96	2.54	8.77
NL/HBO	4.89	5.33	8.00	3.00	9.25

	INDIA	INS_CFS	INS_WFS	INT	NEW
NL/SST	3.98	5.43	5.41	6.41	4.76
NL/PUP	6.80	5.20	5.33	6.50	6.40
NL/USS	4.36	5.3	5.78	7.28	5.37
NL/USL	2.88	5.73	6.08	6.92	4.49
NL/HBO	2.00	6.67	6.80	8.00	3.80

	OFF	PERFECT	PULL	RED	SECONDAR
NL/SST	5.94	8.47	6.70	7.29	4.96
NL/PUP	4.80	9.20	8.80	5.50	5.20
NL/USS	6.54	8.48	7.85	6.85	5.60
NL/USL	8.38	8.82	7.69	7.80	4.16
NL/HBO	8.20	8.80	8.60	9.00	4.80

	SUIT	SURE	THAT	THAT_MAN	THIN
NL/SST	3.12	6.88	7.20	4.26	7.67
NL/PUP	2.40	8.40	7.20	7.00	7.60
NL/USS	2.68	6.88	8.72	3.52	8.78
NL/USL	2.55	4.86	8.12	3.08	8.71
NL/HBO	2.22	3.40	8.20	4.80	8.40

	THOMAS	TIN	TO_WALES	VAN	WINE
NL/SST	5.16	5.24	6.96	7.94	7.73
NL/PUP	5.60	6.00	6.40	7.60	7.00
NL/USS	5.50	7.03	7.07	8.94	8.27
NL/USL	4.31	6.64	6.60	9.04	9.00
NL/HBO	5.40	7.00	7.20	8.80	8.80

In the interest of clarity, and in order to make the ESIs as representative as possible, it is desirable to pool the results of smaller groups of respondents (NL/PUP and NL/HBO) with larger groups whose backgrounds are most similar to them (NL/USS and NL/USL respectively). This can only be done if the ESIs of the relevant groups are also sufficiently similar. A pairwise t-test for the NL/PUP and NL/SST indeed showed that the two groups' indices were not significantly different ($t = -0.916$, $df = 39$, n.s.) and a similar result was obtained for NL/USL and NL/HBO ($t = -1.334$, $df = 39$, n.s.). This implies that the ESIs for NL/PUP and NL/SST can indeed be merged into one single group: NL/STU. Correspondingly, the ESIs for NL/USL and NL/HBO may be collapsed into another single group: NL/LEC.

In order to determine which errors were considered the most severe, the ESIs of all errors were calculated for the remaining three groups (NL/SST, NL/STU and NL/LEC) and ranked from highest to lowest for each error

category. Subsequently, the highest indices in a particular category were identified by selecting those which ranked above the median (or middle value) in at least one of the three groups of respondents. The medians for these groups are listed in Table 2.19.

Table 2.19. Medians of Error Severity Indices per error category and per group of respondents.

Error category	Groups of respondents		
	NL/SST	NL/STU	NL/LEC
Phonemic	6.88	7.00	8.16
Realisational	5.24	6.79	6.63
Distributional	3.98	3.84	2.74
Stress	7.93	7.99	7.54
Suprasegmental	5.18	5.44	5.01

Table 2.20 shows ESIs for all **phonemic** errors, ranked from highest to lowest for the three main groups (NL/SST, NL/STU, NL/LEC). It is immediately apparent that these three rankings are substantially different with only the highly salient error in BED and the two potential distractors THOMAS and EXAM occupying the same position in all three groups. However, if only those errors are considered which have indices above the median in at least one of the three groups, one arrives at a list which is virtually identical for NL/SST, NL/STU and NL/LEC, and includes BED, VAN, WINE, THIN, BAT, THAT and OFF. The above will form the basis for the phonemic errors selected in the follow-up experiment discussed in the remaining chapters.

It is also interesting to note that the ESIs in the phonemic category are relatively high, possibly arguing in favour of including additional tokens from this category in the core experiment. However, most of the phonemic errors in the lower range need not be considered for inclusion, since they are either variations on those already included (such as ANNIE), potential distractors (THOMAS and EXAM), or errors that most native speakers will consider to be unimportant (such as SURE). The fact that SURE was considered to be so very significant by the secondary school teachers, students and pupils may, in this particular case, even be an indication of obsolescent pronunciation standards (see 2.1.3). However, there were two tokens that were assessed surprisingly leniently in all three groups: COLOUR and PULL. Since these are described by Collins & Mees (2003b: 97, 290) as “persistent” errors (see also 2.1.3), they are likely to be judged more severely by native speakers, and may therefore produce interesting results if included in the core experiment.

Table 2.20. Ranking of Error Severity Indices for all phonemic errors, arranged according to the three main group of respondents. Grey shading represents correspondences in ranking between all three groups.

Ranking	NL/SST		NL/STU		NL/LEC	
	Error	ESI	Error	ESI	Error	ESI
(1)	BED	8.31	BED	9.09	BED	9.29
(2)	VAN	7.94	BAT	8.89	BAT	9.19
(3)	WINE	7.73	VAN	8.87	VAN	9.00
(4)	THIN	7.67	THIN	8.72	WINE	8.97
(5)	BAT	7.67	THAT	8.65	THIN	8.66
(6)	THAT	7.20	WINE	8.22	OFF	8.35
(7)	SURE	6.88	PULL	7.00	ANNIE	8.16
(8)	PULL	6.70	ANNIE	6.97	THAT	8.13
(9)	COLOUR	6.04	SURE	6.96	PULL	7.84
(10)	OFF	5.94	OFF	6.45	COLOUR	7.74
(11)	ANNIE	5.81	COLOUR	6.39	SURE	4.62
(12)	THOMAS	5.16	THOMAS	5.51	THOMAS	4.51
(13)	EXAM	3.09	EXAM	3.19	EXAM	2.26

Table 2.21. Ranking of Error Severity Indices for all realisational errors, arranged according to the three main group of respondents. Grey shading represents correspondences in ranking between all three groups.

Ranking	NL/SST		NL/STU		NL/LEC	
	Error	ESI	Error	ESI	Error	ESI
(1)	ICE	6.75	ICE	7.90	RED	8.00
(2)	RED	7.29	DEAD	7.82	ICE	7.97
(3)	DEAD	5.87	TIN	7.02	TIN	6.70
(4)	TIN	5.24	RED	6.79	DEAD	6.63
(5)	HOT	4.96	FULL	5.89	FULL	5.47
(6)	FULL	4.63	HOT	5.28	HOT	4.13
(7)	THAT MAN	4.26	THAT MAN	3.67	THAT MAN	3.37

As Table 2.21 shows, the differences in ranking between the three main groups are much less striking when it comes to **realisational** errors. In at least one of these groups, the tokens ICE, RED, DEAD and TIN have ESIs which are greater than the median. These will be incorporated into the Native-speaker Experiment, unlike two of the remaining errors, which are either potential distractors (THAT_MAN) or unlikely to be viewed as significant by native speakers (HOT). Even though FULL was assessed relatively leniently, this token may produce interesting results in the Native-speaker Experiment, especially in view of the significance ascribed to it by Collins & Mees (2003b: 291), and was therefore included. (See 2.1.3 for a discussion of these and other errors.)

In the category of **distributional** errors (see Table 2.22), the differences in ranking between the three main groups are not very striking either. In fact, three out of eight tokens occupy precisely the same positions, namely the very salient FILM and CAR together with the potential distractor IDEA. With ESIs greater than the median, the first two are to be included in the follow-up experiment, as will be NEW, HOT_TEA and, perhaps somewhat surprisingly, the potential distractor INDIA, which was allocated an index greater than the median in the NL/STU group. (See 2.1.3 for a discussion of the role of IDEA and INDIA as potential distractors in RP and GA.) Since the errors in the lower range are all either distractors or unlikely to be considered serious by native speakers (see 2.1.3), they will not be incorporated into the Native-speaker Experiment.

Table 2.22. Ranking of Error Severity Indices for all distributional errors, arranged according to the three main group of respondents. Grey shading represents correspondences in ranking between all three groups.

	NL/SST		NL/STU		NL/LEC	
Ranking	Error	ESI	Error	ESI	Error	ESI
(1)	FILM	8.02	FILM	8.44	FILM	7.76
(2)	CAR	5.79	CAR	6.18	CAR	4.92
(3)	NEW	4.76	NEW	5.42	HOT_TEA	4.43
(4)	HOT_TEA	4.29	INDIA	4.49	NEW	4.37
(5)	INDIA	3.98	HOT_TEA	3.84	INDIA	2.74
(6)	IDEA	3.17	IDEA	3.11	IDEA	2.61
(7)	SUIT	3.12	FARMER	2.70	SUIT	2.50
(8)	FARMER	2.73	SUIT	2.67	FARMER	2.23

Table 2.23 shows that the rankings of the **stress** errors are also consistently similar across the three groups. The two tokens that have ESIs greater

than the median (IMAGIN and PERFECT) will also be included in the Native-speaker Experiment, unlike the other two, which are either potential distractors (DEVIANT) or unlikely to be assessed severely by most native speakers (ADVERT). In fact, the relatively high indices for ADVERT are remarkable, and may be indicative, at least in this instance, of a strong adherence to prescriptive pronunciation norms in all three Dutch groups (see 2.1.3).

Table 2.23. Ranking of Error Severity Indices for all stress errors, arranged according to the three main group of respondents. Grey shading represents correspondences in ranking between all three groups.

	NL/SST		NL/STU		NL/LEC	
Ranking	Error	ESI	Error	ESI	Error	ESI
(1)	IMAGIN	8.58	PERFECT	8.52	IMAGIN	8.83
(2)	PERFECT	8.47	IMAGIN	8.46	PERFECT	8.82
(3)	ADVERT	7.39	ADVERT	7.52	ADVERT	6.26
(4)	DEVIANT	4.33	DEVIANT	3.48	DEVIANT	3.80

Table 2.24. Ranking of Error Severity Indices for all suprasegmental errors, arranged according to the three main group of respondents. Grey shading represents correspondences in ranking between all three groups.

	NL/SST		NL/STU		NL/LEC	
Ranking	Error	ESI	Error	ESI	Error	ESI
(1)	TO_WALES	6.96	INT	7.24	INT	7.10
(2)	INT	6.41	TO_WALES	7.03	TO_WALES	6.70
(3)	INS_WFS	5.41	INS_WFS	5.77	INS_WFS	6.21
(4)	INS_CFS	5.43	SECONDAR	5.58	INS_CFS	5.87
(5)	SECONDAR	4.96	INS_CFS	5.29	SECONDAR	4.26
(6)	FROM	4.14	FROM	4.90	FROM	3.77
(7)	FREQ_WFS	3.68	FREQ_WFS	4.22	FREQ_WFS	2.62
(8)	FREQ_CFS	2.96	FREQ_CFS	3.51	FREQ_CFS	2.29

As Table 2.24 shows, there are also considerable similarities between the three main groups in the ranking of the **suprasegmental** errors, especially as regards the lower range, which includes FROM, FREQ_WFS and FREQ_CFS.

These distractors need not be considered for incorporation into the follow-up experiment, unlike the other five errors, which all have ESIs which are above the median in at least one of the three groups. The importance attached to intonation is striking – particularly if one considers the very low scores given to the three examples of Dutch intonation provided in the core experiment. One wonders whether the indices for INT would have been similarly high if the Dutch judges had also been presented with actual examples of Dutch intonation patterns, rather than an abstract description of the pronunciation problem in question. This, however, would have been difficult to accomplish in an experiment without sound files, such as was the case in Dutch Experiment.

At this point, it is convenient to summarise which errors are eligible for inclusion in the follow-up experiment discussed in subsequent chapters; see Table 2.25. In addition, it should be noted that none of these include any of the potential distractors or errors that are unlikely to be considered serious by native speakers (see 2.1.3). The only exception is the somewhat dubious token INDIA, which (as presented in the Dutch Experiment) is not a distractor in American English. Care must be taken that this token, once it has been included in the core experiment, serves as a distractor in both RP and GA. This implies that a correctly non-rhotic pronunciation of *windier* should be provided in the RP version (to rhyme with *India*), but that an appropriately rhotic realisation of this word should occur in the GA form. This type of mirroring will also have to be used in case of CAR, NEW, and SECONDAR, except that here the intended error is the use of the most common GA realisation in the RP form and vice versa.

Table 2.25. Errors which are to be included in the Native-speaker Experiment, by error category.

Error category	Errors to be included	Number of tokens
Phonemic	BAT, BED, COLOUR, OFF, PULL, THAT, THIN, VAN, WINE	9
Realisational	DEAD, FULL, ICE, RED, TIN	5
Distributional	CAR, FILM, HOT _ TEA, INDIA, NEW	5
Stress	IMAGIN, PERFECT	2
Suprasegmental	INS _ WFS, INS _ CFS, INT, SECONDAR, TO _ WALES	5
		26

Finally, it should be noted that the Dutch participants did not attach a great deal of importance to errors that are indicative of obsolescent pronunciation norms (see 2.1.3). This implies that the Dutch respondents tended not to refer to antiquated pronunciation models. The only exceptions in this context are SURE and possibly ADVERT. The errors represented by these tokens are much more insignificant than is sometimes assumed.

2.4 The Native-speaker Experiment: design, subjects and procedure

2.4.1 General aims and target groups

The core experiment described in this dissertation was designed to elicit severity evaluations of a number of representative Dutch pronunciation errors in English from as large and diverse a population of native speakers of English as possible. This is in keeping with the general objectives and more practical goals as described in 1.3.1. One of the latter is to determine whether a reliable hierarchy (based on empirical criteria) can be established for different types and categories of error, which can be implemented in English pronunciation training in the Netherlands. A second aim is to discover whether certain errors are prioritised differently by native-speaker judges with dissimilar linguistic backgrounds (such as speakers of RP and GA), and to establish which additional factors play a part in this. As was pointed out in 1.2.3, the assessment of an error may be affected by the indexical factors (e.g. sex or age of the respondents) and by respondents' assessments of the intelligibility and relative appropriateness of the error in the context provided. In addition, it may be influenced by their degree of leniency towards this particular error, or to Dutch-accented English, or to pronunciation errors and standards in general. Furthermore, some judges' evaluations may be affected by the occurrence, in their own or in closely related accents, of common sound realisations that are similar to the Dutch error in question. This will be referred to as "accent similarity". Awareness of such factors may well be important in establishing priorities in pronunciation teaching. A third practical goal is to compare and contrast native-speaker evaluations of representative Dutch pronunciation errors with those of the judges in the Dutch Experiment; this would also be helpful in realigning priorities, where necessary, in pronunciation training in the Netherlands.

The core experiment was set up to meet these objectives in the following ways. Firstly, care was taken to ensure that the relevant pronunciation errors were as representative as possible. This was done by selecting a number of errors from different categories that the three groups of judges in the Dutch Experiment had described as important (see 2.3). The inclusion of these errors also permits a comparison of their evaluation by native speakers with that of the Dutch respondents. In order to enhance the representativeness of the selection, a number of additional errors labelled "significant" or "persistent" in textbooks such as Collins & Mees (2003b) were also incorporated (see 2.3). Some of the errors thus included in the core experiment were similar to realisations that are also common in certain native varieties of English, although this was clearly not the reason for their selection. A spin-off of this added factor was that it enabled the effect of "accent similarity" to be determined (see Chapter 4). The full selection of errors, and details of their presentation in the core experiment, is described in 2.4.3.

In addition, it was considered important to enlist the cooperation of a large and diverse group of native speakers of English for participation in the survey. This was to ensure that the sample population was both sufficiently representative, and diversified enough, to compare and contrast the error evaluations of different subgroups. As was pointed out in 1.2.3, similar research into error hierarchy tends to involve fairly homogeneous groups of native speakers, often either drawn from the UK or the US, but not both. Other possible selection biases involve respondents' levels of education, linguistic naivety, and exposure to Dutch English. Needless to say, a balanced proficiency curriculum should not aim to prepare Dutch learners of English for interactions with one type of native speaker only (see 1.1). Accordingly, the core experiment was designed to be targeted at different groups of native speakers of English from Britain, Ireland, North America and the Antipodes, who are not necessarily highly educated or linguistically sophisticated – or sufficiently familiar with Dutch English to compensate for any pronunciation errors. This meant that participation in the experiment should not presuppose any familiarity with either Dutch or linguistics, or be limited to certain locations (such as the Netherlands or the UK), but that it should be attractive to, and easily accessible for, a wide variety of respondents.

In view of these preconditions, it was decided to employ the format of an Internet survey. Instead of using abstract descriptions of errors involving linguistic terminology, such online questionnaires can be targeted at a broad range of native speakers by incorporating sound files which are presented to them for assessment.¹⁰ If each of these sound files features a single Dutch pronunciation error in an otherwise native-sounding context, respondents should be able to detect and assess the error without needing any knowledge of Dutch or linguistics, provided the native English accent presented in the experiment is one that respondents feel competent to judge. It also presupposes that participants do not experience any technical difficulties when playing these sound files on their computers.¹¹ A more detailed description of the method used, covering each section of the survey, is provided in 2.4.2.

While some may claim that people can only ever accurately judge accents that are very similar to their own, it has been assumed here that for most native speakers, this competence extends to reference accents such as RP and GA. At least, this would seem to be true for those native English-speaking countries

¹⁰ Preston (1999b: 369), however, refers to “some recent language attitude research which has shown that there is little or no difference in evaluations when the stimulus is a category name or an actual speech sample ...”. Nevertheless, while respondents may find it relatively easy to describe their attitudes to certain native-speaker accents, as in the various experiments described by Preston, they will be considerably more challenged by abstract descriptions of foreign-language errors.

¹¹ This is an almost inevitable problem with experiments of this nature. However, care was taken to ensure that respondents' computers did not start up a separate audio player, which could interfere with the experiment (see 2.4.2), and that downloading time was reduced to an acceptable level (see 2.4.3).

where such pronunciation models appear to be sufficiently well-known, either as a result of educational norms or simply by virtue of exposure. For instance, RP may not enjoy the same prestige everywhere, but it is unquestionably a well-known accent in Britain and Ireland, and also, to a lesser extent, in Australia, New Zealand and South Africa. There are even speakers in these Antipodean countries whose accents closely resemble RP (Wells 1982: 301, 594–595, 611, McArthur 2002: 291, 380, 389).¹² Similarly, most North Americans are very familiar with a “mainstream” variety such as GA, or – if the notion is adopted that GA is an imaginary construct (see Preston 2005) – with the closely related accents commonly designated by this term (compare Wells 1982: 470). The principal reason for including RP and GA in the core experiment is also that they are the two pronunciation models most commonly encountered in the Netherlands and, for that matter, the world.

As a result, it was decided that the experiment would be presented in two different versions, each with a different guise: one actor producing Dutch pronunciation errors in an RP context and another doing the same for GA. It is only because of the practical problems involved in finding suitable actors that the number of guises was not expanded to include other native varieties of English. It was assumed that potential respondents who did not feel competent to judge either RP or GA would decline to take part in the experiment, a phenomenon known as “self-selection sampling” (see Bradley 1999: 388). Unfortunately, such self-selection is likely to involve a significant proportion of native speakers, including many of those speaking varieties of English defined by Kachru (1985) as belonging to the “Outer Circle” (e.g. the Indian sub-continent and West Africa). In spite of these limitations, a set-up involving the two main varieties of English that are widely known and taught would still make the core experiment considerably more representative than many similar experiments of this nature. It should also be capable of generating considerable response from judges in Kachru’s (1985) “Inner Circle”, on both sides of the Atlantic and the Pacific, provided it is attractively designed, user-friendly and does not require any linguistic sophistication on the part of the participants. How this was attempted is discussed in 2.4.2.

In order to determine whether factors such as respondents’ sex, age or linguistic background affect their assessments, the experiment included a section in which these biographical data were provided by participants. While it was assumed that respondents stated their sex or age accurately, this is not necessarily true of their attempts to self-identify their accents. Since this is what Preston (1999b: 360) describes as a “linguistic fact (i.e., linguistic objects as viewed by non-linguists)”, it may be represented “accurately, partially accurately, or completely inaccurately” by what he refers to as “folk respondents” (1999: 360). Because of the well-documented mismatches between

¹² Even though the term “Antipodean” is most commonly associated with Australia and New Zealand, in this dissertation it has been used as a cover term to refer to all countries in the Southern hemisphere where English is spoken as a first language, including South Africa.

dialectologists' descriptions of dialect regions and non-linguists' perception of these (see for instance Lance 1999: 313), it was decided not to ask respondents to identify their own accent from a number of pre-selected options. Such an approach would also have encouraged participants to provide socially desirable answers, or to categorise their speech inaccurately because the relevant label was not included in the experiment. For these reasons, respondents were not provided with a limited set of possible labels (even though these would have been easier to process as data), but were presented instead with an open question where they were asked to identify their own accents themselves. How these were subsequently categorised is discussed in 2.5.1.

With a view to collecting information about respondents' attitudes to particular errors, a designated space was available on the online survey form to provide for individual comments on each token. This also gave participants the option of reporting any technical problems they might have encountered while listening to the speech sample in question, or when identifying and assessing the error in question. These textboxes, in addition to a hypertext link to the researcher's email address, also enabled respondents to provide more general comments on pronunciation errors and standards, or on the nature of the experiment. The data collected from the textboxes are discussed and analysed in 3.4.3.

In order to target the online survey at as wide a range of native speakers as possible, it was designed as an "open-web questionnaire" (freely accessible, i.e. neither protected by a password, nor triggered by a mechanism, as in the case of a pop-up survey; see Bradley 1999: 390). Visitors were directed to the questionnaire by means of a so-called "URL-embedded" e-mail inviting them to participate and offering them to take part in a lottery for a small prize as an incentive (Bradley 1999: 392; see also Gunn 2002). The combination of an e-mail "cover letter" with a web-based survey has become a common technique and is described by Solomon (2001: par. 3) as "an especially effective and efficient approach to Internet surveying". The e-mails were sent from the researcher's Utrecht University e-mail address to other electronic addresses in the UK, Ireland, North America and the Antipodes. Many of those addressed were connected to universities and colleges in these countries, as students or staff, and had received the researcher's e-mail through the help of university administrators or colleagues, or had been presented with it as a posting on their departmental or university "listserv", or on a more general mailing list such as the Linguist List.¹³ A few e-mails were also sent to native speakers of English living in non-English-speaking countries such as France and the Netherlands, including those working for the Universities of Leiden and Utrecht. Other addressees originated from randomly selected university websites in native English-speaking countries featuring, for instance, lists of academic experts in different fields, or students' individual homepages. Some addressees will have had the e-mail forwarded to them because it contained the request to pass it on to "any friends, relatives or colleagues that may be interested in this experiment."

¹³ This was posted at www.ling.ed.ac.uk/linguist/issues/13/13-1635.html#2.

This was intended to create the effect of “snowballing”, a technique which, according to Clayton (2004: 3.1, par. 2), may decrease “the reliance on ... voluntary participation” and “the magnitude of the sampling error”.

It may be argued that the primary focus on targeting the academic community and their relations is one of the significant sampling biases involved in this experiment. This was not motivated by a desire to exclude respondents with other educational backgrounds, but by practical considerations such as the researcher’s contacts with, and relatively easy access to, other university networks. As Gunn (2002: par. 15) has pointed out, however, while a number of communities do not enjoy full or partial Internet access, some university campuses are among those “where connectivity is almost universal”, which makes “sample bias with Web surveys not as great a concern in those populations” (Gunn 2002: par. 15; see also Solomon 2001). This is clearly an asset to Internet researchers, and explains why “Web surveys are a more common survey method on university campuses than with the general population” (Gunn 2002: par.15; see also Couper 2000). In other words, academic communities are a well-established target population in Internet surveys.

Needless to say, this does not address the issue of representativeness for the population as a whole – but this problem is in fact inherent in all surveys on the World Wide Web (Couper 2000: 467). A number of attempts were made to redress the balance as much as possible between academic participants and others. In the first instance, it was hoped that snowballing would also help to generate non-academic respondents. In addition, calls to participate in the experiment were also posted on a number of websites dedicated to teaching, local culture, genealogy, expatriate communities and the media. These were intended to appeal both to special interest groups and to the public at large. In addition to online discussion forums (e.g. educationtalk.guardian.co.uk) and Yahoo groups (including the now defunct “ExpatsinHolland”), these included Usenet groups such as alt.usage.english, soc.culture.welsh, soc.culture.scottish, soc.culture.south-africa and nz.general. Online postings (such as Fraser 2002 and Glenallan 2002) show that this did actually prompt some readers to participate in the survey. Interest in the experiment from outside the academic community is also evident from the fact that it was reviewed in a New Zealand computer magazine entitled *Computerworld* (Broatch 2002). In the interest of keeping the experiment as short as possible and therefore less time-consuming, respondents were not asked how they had heard of the experiment (also known as a “tracking question”, see Gaddis 1998) or what their educational background was, and as a result it cannot be established precisely how many respondents were from outside the academic community. However, the personal details volunteered in participants’ e-mails sent to the researcher also suggested that the sample population was not restricted to staff and students of universities and colleges.

After a brief pilot conducted in May 2002, in which the responses of 12 participants from different parts of the English-speaking world were collected, the experiment was available online, from June to September 2002,

from the researcher's own Utrecht University website (www.let.uu.nl/~rias.vandendoel/personal/pronexp/). This resulted in a large number of responses. Submission records show that 577 of 994 people who started the experiment actually completed it, i.e. 58%; this figure does not include six multiple submissions. While 343 single participants completed the British English version of the experiment (i.e. 59% of the total), there were 234 single respondents who finished the American version (41%). More details about respondents' linguistic background (based on their accent self-identifications) will be provided in 2.5.

2.4.2 Sections included in the survey

The survey consisted of five main sections: (1) introduction; (2) "check-in"; (3) instructions and demo; (4) main body of the survey; (5) self-assessment and completion. Each of these sections will be discussed below. For ease of reference, a fully functional copy of the survey has been posted on www.let.uu.nl/~rias.vandendoel/personal/wwstim/pronexpdemo.html/.

The introduction consisted of a web-page designed to be "motivational" (Dillman *et al.* 1998: 7) and to provide information about the survey as recommended by Gaddis (1998): (i) subject, purpose and target group; (ii) estimated time required for completion; (iii) a reassurance that personal information will be treated confidentially; and (iv) the name of the organisation responsible for the survey or under whose auspices it is conducted, and contact details for the researcher. This included a reference to the English Department and the Research Institute for Language and Speech of Utrecht University, the university's logo, and a link to the researcher's home page and e-mail address. These details were provided in an attempt to enlist the co-operation of as many eligible respondents as possible (see also Gunn 2002a). For the same reasons, mention was made of an "incentive" (cf. Gaddis 1998): participants were told they could take part in a lottery for a small prize (this consisted of digital gift certificates redeemable at a well-known online bookstore). The emphasis on the requirement for participants to be native speakers of English was intended to prepare respondents for the "screening question" (Gaddis 1998) in the "check-in" section. In keeping with recommendations made by Dillman *et al.* (1998: 3), the design of the introduction, as indeed that of the rest of the survey, was kept deliberately "plain", so as not to discourage respondents whose computers or browsers could experience technical problems with a "fancy" (i.e. an elaborate) design (see also Gunn 2002).¹⁴

The remaining four parts of the survey were especially designed by a Utrecht University software developer according to the researcher's specifications, and are based on a CGI script developed for web-based surveys known as WWStim (Veenker 2003). This system may be described as presenting "predefined sequences of template based HTML pages", as a result of which a

¹⁴ The default font was set at Verdana (10 point), which had been "designed specifically for the computer screen" and has been found to be the "most preferred ... font at this size" for use on the computer (Bernard *et al.* 2002: par. 12).

survey can be set up “using only one or two templates” supplied by the researchers themselves (Veenker 2003: par.1). In addition, a “stimulus list provided by the experimenter controls which template must be used for a certain page, which codes in that template must be substituted by which words or HTML fragments, and also which data should be recorded (to a results file) when the subject responds to a stimulus” (Veenker 2003: par. 1). While the templates and stimulus lists were indeed provided by the researcher, this particular survey’s special interactive features were added by the software developer.

An indispensable part of a WWStim-based experiment is the “check-in” page in which respondents provide indexical data about themselves that is stored in the subjects’ database (Veenker 2003). In the present survey, this was preceded by a so-called “pre-check-in”, allowing respondents to verify, by clicking on a link to a sound file, whether they could take part in the experiment without their computers starting up a separate audio player – which could interfere with the experiment.¹⁵ Those respondents who did not encounter any problems were linked through to the check-in page, in which they were invited to state their sex and age and to self-identify their accents. In order to encourage participants’ self-identifications, examples of a wide variety of descriptive labels for accents had been provided in an adjacent textbox.¹⁶

In addition, the check-in page also contained a “screening question” (Gaddis 1998) asking respondents to select the radio button marked “Yes” if English was their first language. In order to help define what was intended by the term “first language”, an adjacent textbox was provided in which respondents brought up in non-English speaking countries were encouraged to select “Yes” if they had spoken English all their lives and spoke it totally fluently. If they opted for “No”, respondents were linked through to a “warning” page. This page also appeared when other required information (such as sex, age and accent) had not been provided. The warning page reminded participants that the experiment was intended for native speakers only and that all required fields had to be completed according to the instructions. Apart from preventing incomplete submissions, this procedure was set up to discourage non-native speakers of English from taking part in the experiment. Needless to say, in a survey of this nature, the researcher does not have full control over such a process of self-selection (see Bradley 1999: 388). Consequently, the possibility cannot be excluded that respondents with an L1 other than English also participated in the survey.

¹⁵ This was singled out as an “appealing feature of the site” in *Computerworld*’s review of the experiment (see Broatch 2002: par. 11).

¹⁶ Technical or potentially loaded terms such as “RP” or “GA” had been avoided, although one respondent from Leeds in the UK made it clear in a separate e-mail that she objected to the term “Standard Southern British English”, because she felt it incorrectly implied that a Northerner could not be a speaker of Standard English. In retrospect, it should perhaps have been made clearer that the term “Standard Southern British English” had been intended to refer to one possible standard variety in Britain, rather than being presented as the only one.

Participants were not compelled to provide their e-mail addresses, as this would only be required if they opted to participate in the lottery.¹⁷ Respondents were, however, obliged to make a choice between two versions of the experiment entitled “British English” and “American English”. These were described as being based on “Standard Southern British English” and “Standard American English” respectively, and participants were invited to select the variety appropriate to them. They were explicitly told that if they did not speak either of these varieties (“for instance, if you’re Irish or Australian”), they should select the one they felt “most competent to judge” (see 2.4.1 for a discussion of this). An overview of the versions selected by different groups of participants has been provided in 2.5. Respondents were also alerted to the fact that by selecting a particular version, they would start up the demo and would hear a “trial sentence”.

Apart from additional instructions and/or clarifications, the “instructions and demo” page contained the same components as of any of the 32 pages in the main body of the survey. These components included (i) a visual representation of the relevant audio stimulus that was opened and played when the page was accessed, including a “Repeat” button for repetition of this stimulus; (ii) a “Yes/No” question about the stimulus (“Does this sentence contain a clearly detectable error?”) with further instructions on how to proceed; (iii) a question asking respondents about the nature of the error (if selected as such), featuring (a) hypertext links to Java script pop-up windows providing definitions of the terms used and (b) an interactive version of the sentence allowing participants to locate the position of the error (if defined as segmental); (iv) a multiple-choice question about the seriousness of the error; (v) a textbox for optional additional comments; and (vi) a button allowing respondents to continue to the next page. Figure 2.1 shows how these different components were presented on a sample page from the main body of the survey.

In addition to advising respondents on the nature of the task, the functions of the different buttons and textboxes and the procedure, the Instructions and demo page alerted participants to the presence of possible distractors and stated the number of stimuli included.¹⁸ Furthermore, respondents were also instructed to judge the different errors by the standards of the relevant reference accent. For instance, the British version of the experiment contained the following instruction: “Please note: You may hear pronunciations that you wouldn’t use yourself but that you recognize as being authentic for Standard Southern British English. Do not regard them as errors.” This was intended to provide clarity for those respondents who did not speak either RP or GA and would possibly be

¹⁷ Another reason for this strategy was the finding that “a relatively high percentage of potential respondents stopped completing the surveys ... when asked to supply their email address” (Solomon 2001: par. 10; see also Gunn 2002).

¹⁸ The latter was included to help respondents monitor their progress as they proceeded through the consecutively numbered pages of the survey. This is similar to the “progress bar” recommended by Dillman *et al.* (1998: 13).

deterred from continuing the experiment by instructions that failed to acknowledge their special position as judges of these reference accents. In addition, it was assumed that any effects of “accent similarity” (see 2.4.1) would be all the more significant if respondents were explicitly instructed *not* to factor this into their assessment.

Repeat **You have no authority over any of us.**

Does this sentence contain a clearly detectable error?

No -> please continue with the next sentence

Yes -> please select the error below

Y o u h a v e n o a u t h o r i t y o v e r a n y o f u s .

Pronunciation of [] in the word []

Word stress

Sentence intonation

How serious is this error?

Very serious Serious Not very serious Relatively unimportant

Space for extra comments:

Next sentence Page 3 of 32

Figure 2.1. A sample page from the main body of the survey.

In the introduction and demo, instructions were also provided on how to identify an error as either segmental or supra-segmental (part iii). As in the main body of the survey, the three options (pronunciation, word stress or sentence intonation) were provided, for the benefit of non-linguists, with definitions which appeared in pop-up windows if respondents clicked on the hyperlinked term. Whereas the supra-segmental phenomena were defined by means of simple descriptions (“Word stress refers to the stress in individual words, e.g. saying muSIC instead of MUsic.” and “Sentence intonation refers to the rise and fall of the voice over the complete sentence.”), pronunciation was presented as a residual category (“This includes all speech-related errors other than word stress or sentence intonation”). While all three terms could be selected by clicking on the relevant radio button, it was also possible to opt for “pronunciation” by clicking on the relevant segment in the highlighted interactive sentence directly above (also required for those who had used the radio button to select “pronunciation”). In these highlighted sentences, hyperlinked letters or letter combinations were used to represent the different phonemes contained in the stimulus. (Silent letters had naturally been excluded.) By clicking on these, without having any knowledge of phonetic transcription, or indeed of linguistics,

respondents could select the phoneme which they thought contained the error, and their selection would be displayed automatically under the heading “Pronunciation”. This was explained in detail in the instructions and demo page, and respondents were provided with a highlighted, clickable sentence with which to practise (see Figure 2.2). The arguably very salient error of replacing /ŋ/ by /ŋk/ in the word *feeling*, featured in the stimulus played when the Introduction and demo page was accessed and mentioned in the Instructions and demo, had been selected as a first example, as this would “be easily comprehended and answered by all respondents” (Dillman *et al.* 1998: 8). This error had the added advantage of being rare in the English of Dutch learners (Collins & Mees 2003b: 168), as a result of which it would be unlikely to bias respondents towards any actual pronunciation errors found in this learner variety. (Details of how the stimuli were recorded are provided in 2.4.3.)



Figure 2.2. The interactive sentence used in the “Instructions and demo” section of the survey. Note that all letters and letter combination separated by spaces were individually hyperlinked, with the exception of the silent e in *leave*.

Subsequently, respondents were instructed to continue to part (iv), the multiple-choice question on error severity. Four radio buttons had been provided to help identify this error as either “very serious”, “serious”, “not very serious” or “relatively unimportant”. Together with the option “No” from part (ii), this effectively constituted a 5-point Likert scale ranging from “no error” to “a very serious error” (see Likert 1932). A similar scale had been used in the Dutch Experiment, as a consequence of which it was possible to compare respondents’ assessment of error gravity in the two experiments. No attempt was made to define “serious” or “unimportant” in terms of the different potential effects that various segmental or non-segmental errors may have (such as “unintelligible” or “distracting”), as this could prejudice participants against particular types of error, and would fail to do justice to respondents’ own reasons for assigning importance to these.

In keeping with the recommendations made in Dillman *et al.* (1998: 7–11), attempts were made to indicate clearly, both in the instructions and elsewhere, how respondents were intended to continue from one step to the next, and from one page to the next. This was done by a combination of instructions,

arrows, pop-up windows and buttons.¹⁹ It was hoped that measures such as these would make the survey more user-friendly. Similarly, the menu and toolbars had been disabled on all pages of the survey except the Introduction. This was intended to help guide respondents through the experiment by discouraging the use of “Back” or “Forward” buttons in their different browsers. Apart from creating confusion, using such options could lead to multiple submissions of individual items.²⁰

After respondents’ attention had been drawn to the “Space for extra comments” in (v), they were invited to start the main body of the survey by clicking on a button entitled “BEGIN” (vi). This started up the first of the 32 similar pages generated by WWStim on the basis of the same template. Each of these contained different stimuli drawn from the two stimulus lists (one for RP and one for GA), and showed the corresponding visual and interactive representations of the relevant sentence. All pages were presented in a random order to compensate for any learning effects, and respondents were required to complete all 32 pages in the order provided by the system. It should be noted that this goes against the recommendations made by Dillman *et al.* (1998: 11–12), although they allow exceptions for what they refer to as “order effects” (which presumably also include “learning effects”). The full list of stimuli is discussed in 2.4.3.

Respondents who had assessed all 32 stimuli were subsequently directed to a page containing a final question. This was “How would you describe yourself as a judge of pronunciation?” Participants were presented with a choice of five radio buttons, one of which could be selected to indicate their answer. These options represent five points on a Likert scale and consisted of the following: (1) very lenient, (2) lenient, (3) neither lenient nor strict, (4) strict, and (5) very strict. These data were collected so as to normalise respondents’ assessments for their self-reported leniency, which varies between subjects. As this was a “personal” question, it was not presented until the end of experiment, as is recommended by Frary (1996); see also Gunn (2002). In addition, it was assumed that respondents would find it easier to assess their own leniency once they had actually completed the survey.

¹⁹ For instance, respondents were instructed to skip parts (iii) and (iv) if they had answered “No” in (ii), but to proceed to these sections if they had answered “Yes”. If the latter failed to complete these sections, they were shown a pop-up window reminding them to do so. Another example is that the option “Yes” in (ii) was automatically selected if (iii) had been completed, whether or not respondents had already answered (ii) themselves. In addition, the buttons which allowed respondents to continue to the next page had been clearly marked to indicate this.

²⁰ Nevertheless, 48 respondents, divided roughly equally over the two main versions, managed to produce a total number of 74 double submissions, presumably by using the right-click function of their mice to go back to the previous page. As the second submission (which was the one selected in the data analysis) was only different in eight of these 74 cases, this was not a significant problem.

Only if respondents answered the question about their leniency were they presented with the last page of the experiment, in which they were thanked for their participation, reminded about the lottery and presented with the researcher's university e-mail address in case they had further questions or comments. It is not until this page was accessed that a participant's individual results file was completed, and an e-mail was sent to the researcher. The data obtained could then be processed and combined with other respondents' results files into a larger database for statistical analysis.

2.4.3 Audio stimuli used in the Native-speaker Experiment

In each of the two versions of the Native-speaker Experiment (RP or GA), participants were presented with a total number of 32 audio stimuli (in addition to the demo). Each stimulus consisted of a carrier sentence which contained a single Dutch pronunciation error but which was otherwise no different from an unmarked native-speaker realisation. With the exception of a single distractor, which did not contain any deviation from native-speaker English, the errors in question were either **phonemic**, **realisational**, **distributional**, **stress-related** or **suprasegmental**; for a discussion of the relevant categories, see 2.1.3. A short description of these errors (with their categorisation and a reference to the discussion in Collins & Mees 2003b), is provided in Table 2.26, together with the context (word or phrase) and carrier sentence in which they were presented in the experiment. For the sake of convenience, the errors have been identified by key words (in SMALL CAPITALS) which are as similar to these contexts as possible. For instance, the token labelled COLOUR represents /ʌ ~ ɒ/ confusion in the word *colour*. Table 2.26 also shows the difference between the normal context of the error (transcribed phonemically) and the manipulated context in the stimulus material (transcribed phonetically where relevant). While the phonetic transcriptions show the common substitutions made by Dutch learners of English (e.g. [ɔ] for /ʌ/ in *colour*), the effect of such errors has been described largely in phonemic terms (e.g. /ʌ ~ ɒ/ confusion).

The selection was based on the 26 errors from the Dutch survey presented in Table 2.25 (see 2.3). Where possible, the same words or phrases used to illustrate the error in the Dutch Experiment were also used in the carrier sentences provided in the Native-speaker Experiment (except that CAR, INDIA, NEW, and SECONDAR were “mirrored” in the GA version; see 2.1.3 and 2.3). This was done in the interest of facilitating comparison between the assessments of these errors by the Dutch participants and those by native speakers (see 3.6). However, in three cases a slightly different context was used to exemplify the same error. Whereas in the Dutch Experiment, the minimal pair used to illustrate the conflation of /ʊ/ and /u:/ was *pull* ~ *pool* (as illustrated by the key word PULL), the core experiment instead used the context *stood* (pronounced with [u] rather than /ʊ/) as designated by the key word STOOD. Similarly, while the keywords FULL (in the Dutch Experiment) and TELL (in the core experiment)

both refer to the Dutch tendency to use over-dark [ɫ^s] in English, in one survey this was exemplified by *full* and in the other by *tell*.²¹ Likewise, TIN and TIE both refer to the lack of aspiration associated with word-initial plosives in Dutch English.²² In spite of the slight differences between the various contexts, STOOD, TELL and TIE will be treated as referring to the same errors as PULL, FULL and TIN respectively.

As the Dutch Experiment relied on verbal description of the errors, it was assumed that participants would not necessarily benefit from any attempts to exemplify suprasegmental phenomena using particular words or phrases, so errors in this category were mostly described in more general terms. Needless to say, this method could not be employed in the core experiment, consisting of audio stimuli, and consequently, carrier sentences with realistic examples of suprasegmental errors were used instead. This meant that the abstract error of “insufficient use of weak forms” (INS_WFS) was replaced by a specific example of a failure to use a weak form in the subordinating conjunction *that* (THAT_THA). In addition, a similar example of this phenomenon had already been included in the experiment (TO_WALES).

Likewise, the general error of “insufficient use of contracted forms” (INS_CFS) was substituted for an actual occurrence of this in the modal *would* (WOULD_ON). Furthermore, three concrete examples of intonational deviation were provided (INT1, INT2, INT3), rather than the more broadly phrased error of “too little variation in intonation”. The nature of these intonation errors, and the slightly different method used in recording the relevant carrier sentences, is described at the end of this section.

Table 2.26 (over page). Overview of audio stimuli used in Native-speaker Experiment, headed by key word, followed by description of relevant error and error category; page references are to Collins & Mees (2003b). Note that corresponding carrier sentences and contexts are shown in bold; segments containing errors are underlined. Phonemic transcriptions (based on Wells 2000 for English, and Collins & Mees 2003b for Dutch) indicate differences between normal and stimulus context; phonetic transcriptions (in square brackets) show changes made to stimulus context.

²¹ It should be noted that *pull* was replaced by *stood* as a context so as to avoid any influence of dark-l on the preceding back vowel; similarly, to avoid the vocalisation of dark-l by a close back vowel, *full* was replaced by *tell*.

²² The context *tin* was changed to *tie*. *Tin* is a high-frequency item in British English (and Antipodean varieties) in the sense of “metal container for food, drink and other substances”. In American English its use is largely restricted to the common name for the metal *Stannum*, which would have rendered it difficult to produce a convincing carrier sentence.

	Short description of the error	Error category (+ page ref.)	Carrier sentence and context of intended error	Normal context	Stimulus context
1	BED Final fortis/lenis neutralisation	Phonemic (48–55, 290)	She lay in bed for most of the day.	bed	be[t]
2	BAT /æ ~ e/ confusion	Phonemic (94, 290)	Hundreds of bats fluttered about in the cave.	bæts	b ɛ ts
3	VAN Initial fortis/lenis neutralisation	Phonemic (48–55, 290)	A small blue van was parked across the street.	væn	[f]æən
4	WINE /v ~ w/ confusion	Phonemic (174–175, 290)	They were drinking red wine and eating cheese.	wain	[v]aɪn
5	THIN Substitution of initial /θ/ by /t/	Phonemic (142, 291)	She began to look as thin as a ghost.	θɪn	[t]ɪn
6	AUTHOR Substitution of medial /θ/ by /t/	Phonemic (142, 291)	You have no author ity over any of us.	RP ɔ:ˈθpreɪ GA əθɔ:reɪ	RP ɔ:ˈ[t]preɪ GA ə[t]ɔ:reɪ
7	BOTH Substitution of final /θ/ by /t/	Phonemic (142, 291)	We were both young and inexperienced.	RP beəθ GA booθ	RP beə[t] GA boo[t]
8	OFF Final fortis/lenis neutralisation	Phonemic (48–55, 290)	Many of our students come from English-speaking countries.	əv	ə[f]

9	THAT	Substitution of initial /θ/ by /d/	Phonemic (142, 291)	We were supposed to be meeting that man at two o'clock.	ðæt	[d]æt
10	WEATHER	Substitution of medial /θ/ by /d/	Phonemic (142, 291)	It's unusual to have such cold weather in August.	RP 'weðə GA 'weðɹ	RP 'we[d]ə GA 'we[d]ɹ
11	BREATHE	Substitution of final /θ/ by /d/	Phonemic (142, 291)	The patient began to breathe more regularly.	bri:ð	bri:[d]
12	RED	Use of uvular-r	Realisational (179, 291)	The bus had failed to stop at the red light.	red	RP [r]ed GA [ɹ]ed
13	ICE	Over-long /aɪ/	Realisational (111, 290)	This joke is guaranteed to break the ice at parties.	ais	[a:ɪ]s
14	TIE	Unaspirated [t]	Realisational (150–152, 291)	He always wears a tie in the office.	taɪ	[t̪]aɪ
15	DEAD	Glottalisation of final /d/	Realisational (153, 290)	She had worked so hard she was half dead with exhaustion.	ded	de[ʔ]
16	FILM	Epenthetic [ə] in /lm/	Distributional (171, 291)	We saw a great film on TV last night.	film	'fɪ[ləm]
17	CAR	Inappropriate post-vocalic r	Distributional (180–181, 291)	My friend has a car but I don't think he'll give us a lift.	RP kɑ: GA kɑ:r	RP k[ɑ:ɹ] GA k[ɑ:ɹ]

18	HOT_TEA	Degemination of /#t/	Distributional (218, 290)	In some countries they drink hot tea at four o'clock.	RP hɒt'ti: GA hɑ:ˈtʃi:
19	INDIA	Distractor	Distractor (-)	It's windier here than in the plains of India.	—
20	NEW	Yod deletion/insertion	Distributional (-)	Apparently, Bill's new car has been giving him a lot of trouble.	RP nu: GA nju:
21	IMAGIN	Misplaced stress	Stress (232, 291)	Catherine is one of the more imaginative members of the class.	RP ɪ'mædʒɪnətɪv GA ɪmædʒə'neɪtɪv
22	PERFECT	Misplaced stress	Stress (231, 291)	Fortunately, I've found the perfect solution to this problem.	RP pə'fekt GA p'ɹfekt
23	TO_WALES	Absence of weak form	Suprasegmental (20, 290)	We're going to Wales for a long relaxing holiday.	'tu: weɪlz
24	THAT_THA	Absence of weak form	Suprasegmental (20, 290)	They all said that that may be done very differently.	ðæt ðæt

25	SECONDAR	Absence / presence of weakening	Suprasegmental (-)	The amount of money should really be a secondary consideration.	RP 'sekəndri GA 'sekəndri	RP 'sekəndri GA 'sekəndri
26	WOULD_ ON	Absence of contracted form	Suprasegmental (20, 290)	I'd like to tell her what he's up to, but she would only go and let the cat out of the bag.	RP wəd 'əʊnli GA wəd 'əʊnli	RP 'wɒd əʊnli GA 'wɒd əʊnli
27	TELL	Overdark pharyngealised [H]	Realisational (170–171, 291)	My mother refused to te ll me the truth.	tel	te[f ^h]
28	COLOUR	/ʌ ~ ɒ/ confusion	Phonemic (97, 291)	Actually, my stepfather is totally col our-blind.	RP 'kʌlə GA 'kʌlɹ	RP 'k[ɔ]lə GA 'k[ɔ]lɹ
29	STOOD	/ʊ ~ u:/ confusion	Phonemic (97, 290)	He st ood still for a long time.	stɒd	st[u]d
30	INT1	Intonational deviation	Suprasegmental (291)	They think it's totally stupid.	—	—
31	INT2	Intonational deviation	Suprasegmental (291)	I didn't actually think that was true, but you may be right.	—	—
32	INT3	Intonational deviation	Suprasegmental (291)	Are you taking the car?	—	—

If one also disregards INDIA, which was the only distractor included in the core experiment (on the strength of the importance attached to it by NL/STU respondents), this leaves a set of 22 errors that are similar in both surveys. A more detailed discussion of these may be found in 2.1.3 (as well as in Chapters 3 and 4), while the differences between native and non-native assessments of these errors will be discussed in 3.6. Other numerical discrepancies between the selection of 26 errors in the Dutch survey and the 32 stimuli in the Native-speaker Experiment will be accounted for below (see also Table 2.27).

Table 2.27. Comparison of non-identical errors in the two experiments.

Dutch survey	Native-speaker survey	Comment
FULL	TELL	Same error, different context
PULL	STOOD	
TIN	TIE	
INS_CS	WOULD_ON	Similar but incompatible (general ~ specific)
INS_WFS	THAT_THA	
INT	INT1, INT2, INT3	
–	AUTHOR	Additional errors
–	BOTH	
–	BREATHE	
–	WEATHER	

Apart from the additional intonation errors, the core experiment included four more examples of the substitution of /θ, ð/ by [t, d] in medial position (as in *authority* and *weather*) and in final position (*both* and *breathe*). This was done for a number of reasons. Firstly, while most other segmental errors included in the Native-speaker Experiment are typically found in a particular position in a word or syllable, this is not true of /θ, ð/ substitution. For instance, some errors are found typically, or even exclusively, word-initially (e.g. unaspirated /t/), medially (e.g. degemination) or finally (e.g. glottalisation of final /d/). Certain other errors are restricted to particular positions, for instance those involving r-distribution or checked vowels. Unlike these errors, substitutions of /θ, ð/ by [t, d] may be found in any position of the word, and are widespread in Dutch English (even though Collins & Mees 2003b: 142 maintain that substitutions involving /s, z, f/ are preferred in certain positions). It was therefore decided to increase the number of stimuli to include examples of these substitutions in medial and final position. This would make it possible to discuss the effect of an error's position on its assessment by different groups of native speakers. This is especially relevant in the case of stop realisations of /θ, ð/. While these are also found in different varieties of English (see 4.2.5 and 4.4.5

for a detailed discussion), they are not always equally common in all positions (cf. Wells 1982: 516, Wolfram & Schilling-Estes 1998: 324–325), and this may affect native speakers' evaluations of these. If this is indeed the case, it must be factored into any attempts to establish a hierarchy of error.

It was decided not to increase the number of distractors, so as to keep the experiment as short as possible, and therefore attractive to potential respondents. In addition, it was assumed that all carrier sentences contained a large number of segments and supra-segmental features which could also serve as distractors if they were erroneously identified as deviant. The different hit rates for each intended error testify to this (see 3.4.4). It may be argued that in this respect, the intonation errors should be treated differently from the segmental and other suprasegmental errors, as they affect the entire utterance, and that in each instance an intonationally non-deviant version of the carrier sentence should also have been provided. In retrospect, this could be seen as a fault in the design of the experiment.

Different actors were employed to read out the two versions of the experiment – one for RP and another for GA. In the interest of controlling all other variables, which is a standard requirement of the matched guise technique (Lambert 1967), care was taken to ascertain that the two actors in question had maximally similar backgrounds. They were both bilingual speakers of Dutch and English, male and aged over 55, and both spoke educated standard varieties of English (RP and GA respectively), obtained as a result of extended periods of time spent in native-English-speaking environments. As tenured full-time lecturers at Dutch universities, they had accumulated considerable experience of teaching English pronunciation to Dutch students. Both were accomplished mimics of Dutch pronunciation errors for didactic purposes, and were well-suited to the task of reading out the 32 stimuli in their native English accents while incorporating a single Dutch segment or supra-segmental phenomenon.

The RP actor was recorded in a sound-insulated booth at the University of Utrecht phonetics laboratory, employing a Sennheiser ME 64 unidirectional condenser microphone and a high-quality DAT recorder; the GA actor was recorded under similar conditions at the University of Leiden phonetics laboratory, except that a Sennheiser MKH 416 unidirectional condenser microphone was used. The recordings were digitised (16 kHz, 16 bits) and subsequently edited in the speech processing programmes GIPOS (Vogten & Gigi 2002) and PRAAT (Boersma & Weenink 2002). The actors' performances were carefully checked and approved by two trained phoneticians and also by 12 native speakers of different varieties of English before being used in the experiment.

A somewhat different procedure was followed for the intonation tokens, where three stimulus utterances were produced which contained deliberate deviations from the RP or GA norm along the dimension of speech melody. In order to obtain the intonationally deviant stimuli, a male native speaker of Dutch and near-native speaker of English, who was also a professional phonetician and specialist in intonation, recorded the utterances (1) *They think it's totally stupid*; (2) *I did not actually think that was true but you may be right*; and (3) *Are you*

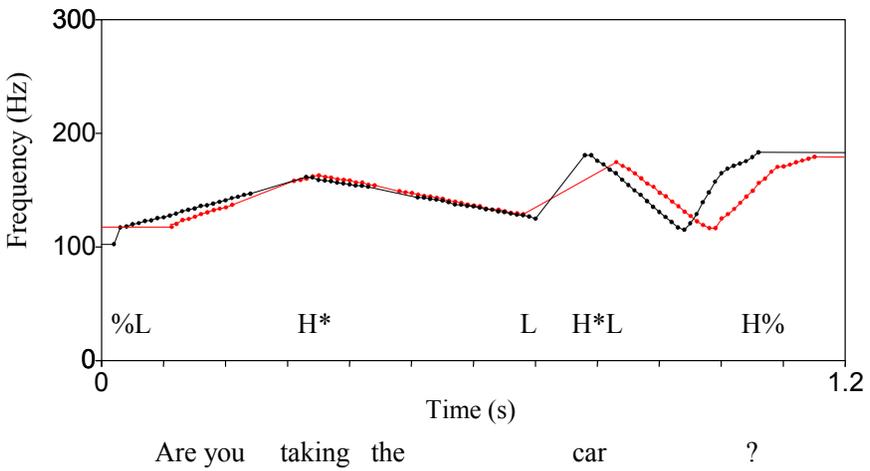
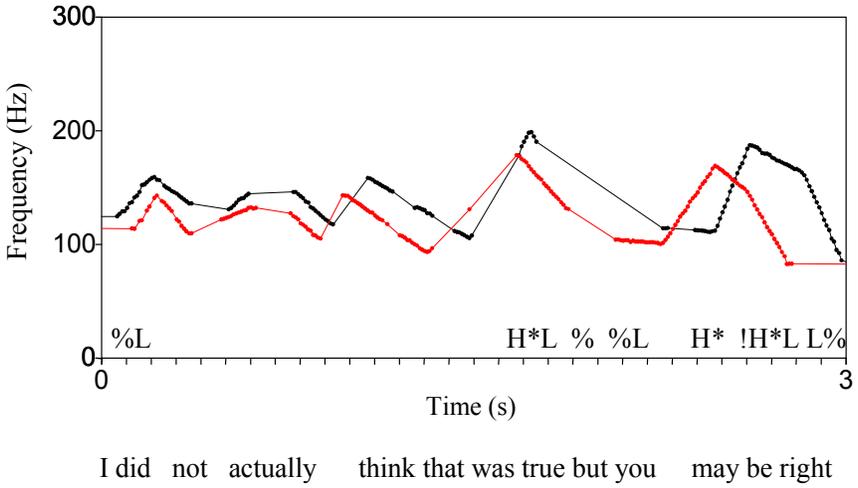
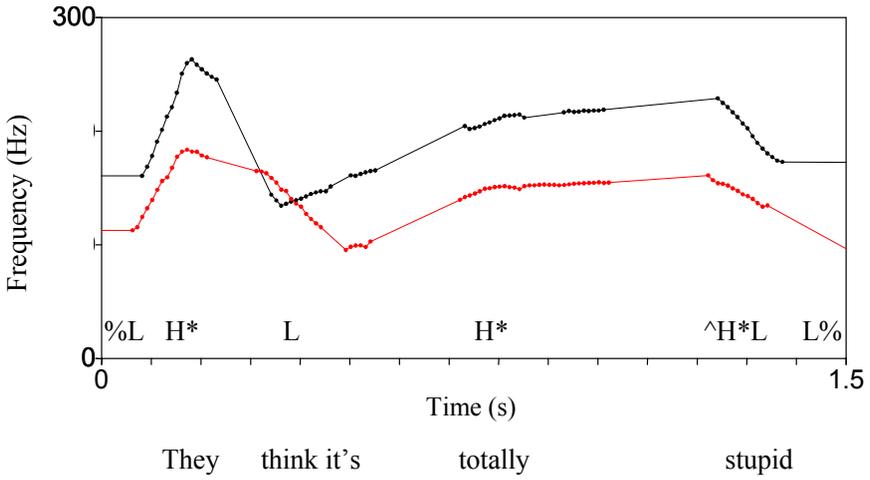
taking the car?, observing near-native British English segmental quality but using typically Dutch intonation patterns.

These recordings were digitised (16 kHz, 16 bits) and the fundamental frequency was extracted using the autocorrelation method implemented in the PRAAT speech processing software. The F_0 curves were then interactively stylised by means of the PSOLA analysis and re-synthesis technique implemented in PRAAT (see also Rietveld & Van Heuven 2001: 379–380) such that they were replaced by the smallest number of straight line segments (in a linear time by $\log F_0$, i.e. semitone, representation) required to generate a melodically equivalent version. The stylised Dutch melodies were then imported into recordings of the same sentences read by the RP and GA actors (under similar conditions as the other 29 carrier sentences). The imported or “transplanted” curves were given approximately the same mean pitch as the original curves they replaced. However, the excursion sizes of the pitch movements in the stylised contours were not affected. In addition, the pivot points in the stylised contours were time-shifted such that their segmental alignment was the same as it was in the utterance spoken by the Dutch phonetician. For instance, if a particular pitch peak occurred at the temporal midpoint of the original syllable, it was given the same relative timing in the hybrid version.

These manipulations ensured that the imported contours would retain all the characteristics of Dutch melodies, such as the characteristically narrower pitch range (see Willems 1982, De Bot 1982, De Pijper 1983) of Dutch, and typically Dutch timing. The mean pitch of the utterances, however, mimics the characteristics of the individual RP or GA actors. The possible effects which these manipulated utterances, as presented in INT1, INT2 and INT3, could have on different groups of native speakers of English have been discussed in detail in 3.5.23.

The resulting pitch curves (after re-synthesis) of the melodically deviant utterances are presented in Figures 2.3, 2.4, and 2.5. The annotations provided in the figures are ToDI transcriptions (Gussenhoven *et al.* 2003, Rietveld & Van Heuven 2001: 399–401) of the melodic pattern aimed at by the Dutch speaker.

Figures 2.3 to 2.5. Pitch curves (stylised) superposed on the RP (black) and GA (grey) stimulus utterances represented by INT1, INT2 and INT3. ToDI labelling is indicated.



Since, at the time of the experiment, relatively slow modem connections were in much more common use than broadband (the latter allowing for virtually instantaneous downloading) it was decided to attempt to reduce downloading time to an acceptable level. In view of this, all 32 sound files were downsampled to 11.025 kHz. Subsequently, file size was further halved by G.711 μ -law encoding using the digital audio editing programme GoldWave (2002). As a result, the average downloading time of a sound file would be no more than roughly 4.5 seconds on a standard 56k bit modem. This would have been doubled if μ -law encoding had not been used.

2.5 The Native-speaker Experiment: data processing

2.5.1 Analysis and categorisation of accent self-identifications

For the reasons discussed in 2.4.1, respondents were provided with an open question where they were asked to describe their own accents. The resulting accent self-identifications were subsequently arranged in two separate lists (for each of the two versions of the Native-speaker Experiment) and silently edited prior to categorisation.²³ The two lists were then analysed to see if certain patterns emerged that would allow classifications into a number of discrete groups with well-established accent characteristics. This meant, for instance, that any respondent who had used the term “Standard” (without further modification such as “Northern”) was placed in a group with other self-styled speakers of the most widely recognised “prestige” varieties: RP or GA. Accent self-identifications using any other well-known linguistic or folk-linguistic labels for these varieties (ranging from “Gimson RP” to “public school” or “the Queen’s English” in the RP version, and from “Standard American” to “US” in the GA version) were also included in these categories. Other accent labels were also chosen to do as much justice as possible to the speakers’ accent self-identifications while still referring to well-documented dialect divisions.

Since respondents were free to choose either version of the experiment (an essential option for those speakers who were neither British nor North American), one American and one Canadian ended up doing the RP version of the experiment, whilst a few respondents from the Antipodes and Ireland – but not a single British informant – opted to do the GA version (see Table 2.28). As these two male North Americans (Subjects 282 and 469) had opted for the British version in spite of the fact that an American version was also available,

²³ This was done to remove spelling errors and inconsistencies in capitalisation and spacing. Punctuation was standardised by replacing all punctuation marks by slashes. Exceptions were made in the case of brackets used to provide additional information such as “(mild)”, or “(expat)”, and in the case of hyphens used for compound adjectives such as “Southern-influenced”.

this raises the question – perhaps more significantly than for any other of the judges – whether or not they assessed the errors from the perspective of RP, or from that of GA. Since this would make it much more difficult to interpret their scores for error detection, error severity and accent similarity reliably, these subjects were excluded from further consideration.²⁴ This also applied to one female Australian and two male New Zealanders (Subjects 385, 745 and 841) who chose the GA version of the experiment. Although some Australians and New Zealanders might perceive their variety of English as having as much affinity with American English as with British English, from a phonological point of view at least, Australian and New Zealand English are considered to be much closer to RP than to GA.²⁵ Irrespective of the reasons that prompted these judges to go for American English, it should be remembered that their judgements of a Dutch learners' pronunciation errors in American English can only be analysed with considerable difficulty.

In addition, certain other respondents were also excluded from further consideration on the basis of their accent self-identifications; a full list is provided in Table 2.28. *Inter alia*, this list takes in hybrid accents, such as “Canadian/British”, “British Southern-Northern mix” or “American/Standard, with some Chicago and Texas features”. The reason for their exclusion was that it was impossible to categorise such respondents for the purposes of the experiment. Any other accent labels that were either ambiguous or referred to insufficiently well-documented varieties of English, such as “Standard Malaysian English”, or “British (East African colonial)”, were similarly removed from consideration.²⁶ This also applied to the male Irish respondent who had labelled his accent “a mix between North-western and Western Irish” (Subject 398), regardless of the well-documented differences between all Southern varieties of Irish English, on the one hand, and those from the North and Northwest, on the other (Hickey 2004: 72–73, 76–80). It should be noted that this respondent was also the only one from Ireland to have opted for the GA version of the experiment.

²⁴ If they had been categorised as belonging to North American accent groups, it would have been difficult to assess their judgements of CAR, NEW and SECONDAR, which are intended to test native speakers' tolerance of RP-GA variety mixing in the English of Dutch learners. It would, of course, be possible to enter missing values for those tokens, or to treat these judgements as North American reactions to Dutch speakers' errors in British English, but this approach would be unlikely to yield clear results.

²⁵ For instance, Wells (1982: 595) states: “Phonologically, all Australian English is very close to RP; phonetically, it is not”. On the close similarities between Australian English, New Zealand English and South African English, see Wells (1982: 592, 605).

²⁶ For instance, the description of Malaysian English provided by Baskaran (2004) makes clear that it is difficult to define its acrolectal variety precisely in phonological terms. Similarly, Schmieid (2004: 921) omits any detailed description of “White African English” from his overview of East African English, as this variety is considered to be “relatively insignificant”.

Table 2.28. Accent self-identifications (in alphabetical order) on the basis of which respondents were excluded from further consideration. Square brackets have been used to indicate that the respondents had decided to do an unexpected version of the experiment.

<ul style="list-style-type: none"> • American / Academic (3 years in England) • American / grew up with West Virginia rural / modified over the years • American / Southern & West Coast • American / Standard (Midwest/West Coast) • American / Standard, with some Chicago and Texas features • American / West [RP version] • A mix between North-western and Western Irish [GA version] • Australian [GA version] • Australian / British born • Both British and American • British (East African colonial) • British / American • British / Middle class London with a Welsh twang • British / Standard / Canadian • British Southern/Northern mix • British Standard and British Northern England • British with American accent 	<ul style="list-style-type: none"> • British / South-North hybrid! • Canadian / British • Canadian / British • Canadian / Toronto [RP version] • English Canadian • Irish / American • New Zealand [GA version] • New Zealand / North Island Urban [GA version] • Scottish / Canadian • Scottish accent with New Zealand accent and Canadian accent all mixed together • Southern Irish with an American accent and tinges of Australian, as I have travelled extensively • Standard American with various influences including Philippine English • Standard Malaysian English • Standard Southern British English with Scottish influences • Western Canadian mixed with Eastern US
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The remaining number of respondents was 545, of which 323 opted for the RP version (i.e. 59%), and 222 for the GA version (41%). All of these were categorised into seven *major* accent groups and, in addition, into 22 *minor* accent groups, the terms *major* and *minor* being employed not to indicate any relative importance, but to allow for a distinction between broad and more specific categories. This made it possible to use the categorisation into *minor* accent groups in those analyses which involve only a few other variables (such as age, sex or leniency), or those where the distinctions are particularly relevant (as with “accent similarity”, see Chapter 4). In those cases, however, where a great many other variables are concerned (such as the token-by-token analysis in 3.5), or where it is useful to make broader generalizations, it was found necessary to refer to *major* accent groups. Table 2.29 lists the seven *major* accent groups, largely categorised by national origin. Note that the special accent

groups created for GA and RP are a consequence of these varieties being used in the experiment. The division into *minor* accent groups will be discussed in the remainder of this section.

Table 2.29. Major accent groups included in the Native-speaker Experiment. Abbreviations used for minor accent groups are explained in Tables 2.30 and 2.36 below.

Major accent group	n	Minor accent groups included	Description of major accent
GB/RP	139	GB/RP	British English - RP
GB/NRP	118	GB/LO; GB/SO; GB/MI; GB/NO; GB/WA; GB/SC; GB/SG	British English - other than RP
IRL	33	IRL/S; IRL/N	Irish English (Northern and Southern)
AU&NZ&SA	33	AU; NZ; SA	Australian, New Zealand and South African English
US/GA	86	US/GA	American English - GA
US/NGA	96	US/EC; US/MW; US/NC; US/NE; US/NY; US/SO; US/WC	American English - other than GA
CDN	40	CDN	Canadian English

The categorisation of respondents taking part in the RP version resulted in 13 minor accent groups (listed in Table 2.30). These categories are all based on respondents' own self-identifications, but they also correspond closely to the labels used in Wells's (1982) authoritative three-volume description of English accents worldwide (any differences from Wells's categorisation will be mentioned in the discussion of these minor accent groups below). In spite of this, it is important to realise that even if respondents' self-identifications are similar or identical to well-known linguistic categorisations, these do not necessarily refer to the same concept. For each minor accent group, a separate table will be provided to show the range of self-identifications used to categorise respondents. Note that in the interest of brevity, identical labels used by different respondents are not repeated.

Table 2.30. Minor accent groups included in the RP version of the experiment.

Minor accent group	n	Description of minor accent
GB/RP	139	British English - RP/Standard Southern
GB/LO	8	British English - Greater London
GB/SO	33	British English - Southern
GB/MI	8	British English - Midlands

GB/NO	48	British English - Northern or Northern-influenced
GB/WA	8	British English - Wales
GB/SC	12	British English - Scots & Scottish
GB/SG	1	British English - Scottish Gaelic English
IRL/N	5	Irish English - North & Northwest
IRL/S	28	Irish English - Southern
AU	7	Australian English
NZ	20	New Zealand English
SA	6	South African English

Table 2.31. Accent self-identifications on the basis of which respondents were placed in the GB/RP major and minor accent group.

<p>British English - RP / Standard Southern</p> <ul style="list-style-type: none"> • BBC English • Brit / Std / Southern • British • British (more or less RP) • British (RP) • British / Public School • British / RP • British / Southern (educated) • British / Standard • British / Standard / Neutral • British / Standard / South • British / Standard / Southern • British / Standard Southern • British / very close to “ideal” RP • British Received Pronunciation • British Standard • British Standard (RP) • British Standard / Southern • British Standard Southern 	<ul style="list-style-type: none"> • British Standard Southern (RP) • British Standard Southern / RP • English • English / Standard • Gimson RP • Proper English (Southern, like the Queen but not as posh) • RP • Standard • Standard British • Standard British / International • Standard British English • Standard British, RP • Standard English • Standard “English” English • Standard Southern British • Standard Southern British English • The Queen’s English • UK / Standard / Southern
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The attempt to be as faithful as possible to judges’ original accent self-identifications has led to the creation of a number of categories that may allow for some overlap. For instance, the category GB/LO or “British English - Greater London” was used for any respondent who specifically mentioned the metropolitan area, while participants were categorised as GB/SO or “British English - Southern” if they employed the terms “Southern”, “South-eastern” or “Home Counties”, but failed to mention either “London” or “Standard”

(see Table 2.32). While the “Greater London” label can only be taken to refer to speakers who would identify their own accents as being on a cline between RP and Cockney, the “Southern” label could theoretically include anyone from Land’s End to Essex. It is certainly conceivable that any accent self-identifications categorised as Southern may still have “Greater London” features – but in the interest of clarity this option has been disregarded in the analysis of the results. Wells (1982), too, maintains a distinction between the accents of “London” and “the south”, but also stresses the influence of London as a “linguistic centre of gravity” on the rest of England (1982: 301), and on adjacent counties in particular (1982: 335).

Table 2.32. Accent self-identifications on the basis of which respondents were placed in the GB/LO, GB/SO or GB/MI minor accent groups respectively.

<p>British English - Greater London</p> <ul style="list-style-type: none"> • British / London • British / London / S.East • British / South London • British London • London • Southern London 	<p>British English - Midlands</p> <ul style="list-style-type: none"> • British / I’m from the Midlands, am used to Northern Standard and Brummie accents • British / Midlands • British / Midlands / Middle class • British Midland • British Northern (Midlands) • British Standard Welsh Border • West Midlands English
<p>British English - Southern</p> <ul style="list-style-type: none"> • British / Home Counties • British / South East • British / Southern • British Southern • South-East British • Southern • Southern (Essex) • Southern British • True England English / Southern 	

Tables 2.32 and 2.33 show which respondents have been categorised as GB/MI (British English - Midlands) and GB/NO (British English - Northern or Northern-influenced). It is likely that the desire to retain accent self-identifications wherever possible may have contributed to some overlap between “Midlands” and “Northern” – although this is inevitable, given the absence of any uncontroversial and coterminous geographical, political and linguistic boundaries (see Beal 2004: 113–115). Wells (1982: 349–350) even includes the Midlands in his discussion of Northern accents. It is only in a few cases where topography seems clearly to assign participants either to the North (as in “Tyne-side”, “Geordie” or “Lancashire”) or to the Midlands (as in “Welsh Border”). In all other cases, respondents’ accent self-identifications as “Northern” or

“Midlands” have been accepted unquestioningly, possibly resulting in participants from the Midlands occasionally labelling themselves “Northern”. In the absence of detailed information on the dialect differences between “Northern English” and “Standard Northern English” (but in recognition of the emergence of a “pan-northern” model, see Beal 2004: 120), these two groups have been taken together.

Table 2.33. Accent self-identifications on the basis of which respondents were placed in the GB/NO minor accent group.

<p>British English - Northern or Northern-influenced</p> <ul style="list-style-type: none"> • British / Lancashire • British / Northeastern • British / Northern • British / Northern (Yorkshire) • British / slightly Northern • British / Standard (Northern-ish) • British / Standard / Northern • British / Standard with a touch of Northern • British / Standard with slight Northern accent • British / Standard with some Northern lexical and phonological features • British North East England (Geordie) 	<ul style="list-style-type: none"> • British North West • British Northern • British Northern English • British Standard / Northern • British Standard, Northern accent • Geordie (mild) • Mancunian (Manchester / Northern) • North East England (Tyneside) • Northern • Northern British • Northern English • Standard British English with a slightly Northern tinge • Teesside English (Educated Northern English) • West Yorkshire English / Standard
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Respondents from other parts of the UK and Ireland have been categorised according to the geographical labels that they themselves provided (see Table 2.34). Whilst, in the interest of conciseness, topographical detail was largely ignored (for instance, “Cork” was placed in the Southern Irish category, and “South Wales” with the rest of Wales), linguistic information that may be expected to impact respondents’ accents strongly was, however, taken into account. This explains the creation of a specific category for English influenced by Scots Gaelic. It also accounts for the inclusion of “Northwestern Irish” with “Irish English - North & Northwest”, as these accents are well-known to be very similar (see, for instance, Wells 1982: 437, Hickey 2004: 72–73). The resulting categories for British English are “Scots & Scottish” (see Wells 1982: 393–412); “Scots Gaelic English” (see “The Highlands and Islands” in Wells 1982: 412–414) and “Wales” (see Wells 1982: 377–393). For Irish English, these are “North & Northwest” and “Southern”. Wells (1982: 417–450) also treats the former separately from the latter. Hickey’s (2004: 73) description of Irish

English also emphasises that the “north of the country is quite distinct from the south, accents of northerners being immediately recognisable to southerners”.

Table 2.34. Accent self-identifications on the basis of which respondents were placed in the GB/SC, GB/SG, GB/WA, IRL/N and IRL/S minor accent groups.

<p>British English - Scots & Scottish</p> <ul style="list-style-type: none"> • British / Scottish • British / Standard / Lowland Scots • North East Scottish (Aberdeen/ Dundee area) but softened in recent years • Scottish • Scottish / English • Scottish English • Standard Scottish English 	<p>Irish English - Southern</p> <ul style="list-style-type: none"> • British Standard (Irish) • Hiberno Irish • Irish • Irish / Dublin • Irish / South • Irish English • Rep. of Ireland • Southern Irish • Southern Irish (expat.) • Southern Irish Cork
<p>British English - Scottish Gaelic English</p> <ul style="list-style-type: none"> • Scottish Gaelic English 	<p>Irish English - North & Northwest</p> <ul style="list-style-type: none"> • British N.I. • Northern Irish • Northern Irish with a bit of Southern • Northwestern Irish
<p>British English - Wales</p> <ul style="list-style-type: none"> • British / South Wales • British / Wales • British / Welsh • English / Welsh 	

Respondents from Australia, New Zealand and South Africa provided no information about their linguistic background other than their national provenance and were consequently categorised as such. The single exception to this was a male New Zealander (Subject 841) who used the label “North Island Urban” to describe his accent, but since this respondent had completed the GA version of the experiment, he has been excluded from further consideration (see Table 2.28). The resulting categories (cf. Wells 1982: 592: 622) are listed in Table 2.35. It should be noted that Wells (1982: 592) also states that it is “appropriate to group these three regional forms under the common heading of southern-hemisphere English”. This is also found in Trudgill & Hannah (2002: 15–30). In the same way, the three “minor” accent groups of Antipodean origin have also been combined in the major accent group “AU&NZ&SA”.

Table 2.35. Accent self-identifications on the basis of which respondents were placed in the AU, NZ and SA minor accent groups.

<p>Australian English</p> <ul style="list-style-type: none"> • Australian 	<p>South African English</p> <ul style="list-style-type: none"> • S.African • South African
<p>New Zealand English</p> <ul style="list-style-type: none"> • British / New Zealand • New Zealand • New Zealand English • NZ 	

Those participating in the GA form of the experiment were divided into nine minor accent groups (listed in Table 2.36), again largely according to respondents' own self-identifications. As will be discussed below (where relevant), these labels are similar to those employed in Wells (1982), and, where Wells does not provide relevant details, to those in Labov (1991) and Trudgill & Hannah (2002). Tables will be provided for each of the minor accent groups, as was done for the RP version, so as to indicate the range of self-identifications used to categorise respondents. Identical labels have not been employed more than once.

Table 2.36. Minor accent groups included in the GA version of the experiment.

Minor accent group	n	Description of minor accent
US/GA	86	American English - Standard American English / General American
US/MW	37	American English - Midwest
US/NC	6	American English - Northern / Northern Cities
US/NE	7	American English - Northeastern
US/NY	5	American English - New York City
US/WC	22	American English - West & Southwest
US/SO	12	American English - Southern
US/EC	7	American English - East Coast
CDN	40	Canadian English

It should be noted that the GA/Standard American English group is identical to the major accent group of the same name. A breakdown of the relevant accent self-identifications is to be found in Table 2.37. As with RP/Standard Southern British, all respondents who labelled themselves “Standard” (without further modification) were placed in this group, together with respondents using any other well-known linguistic or folk-linguistic labels for this type of accent. It is true that the standardness or neutrality of Standard American English or “General American” is debatable – McArthur (2002: 170–171) describes GA as “controversial” – but within the context of accent self-identifications it is convenient to assume that any informants reporting themselves to be speakers of the standard variety aspire to speak according to that particular model, whether this is idealised or in fact a linguistic reality. It may be true that Standard American English is a fiction; see, for instance, Preston’s (2005) discussion of GA as a non-existent learner model, in an article appropriately entitled “How can you learn a language that isn’t there”. However, 47% of US respondents used this “fiction” to describe their own accents.

Table 2.37. Accent self-identifications on the basis of which respondents were placed in the US/GA major and minor accent group.

<p>American English - Standard American English/ General American</p> <ul style="list-style-type: none"> • American • American / mostly Standard • American / Standard • American English • American International • American Standard • American Standard (grew up in South, have lived in Midwest for almost a decade) 	<ul style="list-style-type: none"> • General American • more or less standard American • Standard • Standard American • Standard American English • US • US Standard
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

As with the RP version of the experiment, an attempt has been made to be as faithful as possible to the American and Canadian judges’ original accent self-identifications, at least where the minor accent groups are concerned. Consequently, dialect classifications not employed by participants have been avoided (e.g. “Midland” and “Central Eastern”, as in Trudgill & Hannah 2002: 43–44). In addition, using respondents’ own descriptions has inevitably created some overlap. This is notably true for the labels “Midwest” “Northern” and “Northeastern”. Lance (1999: 313) has demonstrated that the areas constituting dialect regions as perceived by non-linguists are defined very differently by participants from different geographical locations:

If a professor or newsperson refers to Midwestern speech or to the Midwest as a region, we now know that some participants ... will conjure up a map in which Indiana occupies a prominent position, whereas Missouri or Colorado will be prominent in others' mental maps. Ohio may be in one person's Northeastern map and in another's Northern map.

This makes it very difficult to interpret self-identifications such as “Midwestern”, “Northern” or “Northeastern” reliably, and to assign these unequivocally to linguistic classifications. In addition, there is probably considerable overlap between labels such as “Midwestern” and “GA” (see, for instance, Gordon 2004b: 334). In the present study, respondents' own accent labels have simply been retained without further interpretation (see Table 2.38) – apart from the inclusion of a few additional individual participants whose self-identifications placed them more or less unambiguously in one particular category. For instance, the respondent who described her accent as “Upper Midwest” (Subject 585) was placed in the “Midwestern” category, whereas the two respondents who referred to their dialect as “New England” (Subjects 493 and 699) were placed in the “Northeastern” category. No matter how difficult it is to pinpoint the Northeast precisely, few people would argue that it does not include New England (regardless of how narrowly or broadly the latter is defined). The label “Northern / Northern Cities” has been used to categorise all respondents who described their accents either as “Northern” or as “Northern Cities”. This is because it is impossible to ascertain whether or not respondents who refer to their accents as “Northern” actually live in those Northern areas characterised by Labov (1991: 14) as being involved in the Northern Cities Chain Shift (see also Trudgill & Hannah 2002: 45). The accent group also takes in two informants (Subjects 102 and 305) who do not identify their accents in those terms, but since their self-identifications (“Minnesotan” and “upstate New York” respectively) refer to the areas included in the shift by Labov, they have been placed in this category. Since there is no way of knowing, short of interviewing them, whether or not any respondents in this category in fact speak with accents which take part in the Northern Cities Chain Shift, it will be assumed that only a minority of them actually do so.

Table 2.38. Accent self-identifications on the basis of which respondents were placed in the US/NC, US/MW and US/NE minor accent groups.

American English - Northern / Northern Cities

- American / Minnesotan
- American / Northern
- American / upstate NY

- Northern
- Northern Cities American English
- US / Northern

American English - Midwest

- American / Midwest
- American / Midwest (Chicago)
- American / Midwestern
- American / Standard / Midwest
- American Midwest
- American Standard / Midwest
- American, Mid-West
- Midwest American
- Midwestern
- Midwestern American
- Mid-western American
- Mid-Western US
- Standard Midwest American
- Standard Midwestern American
- Upper Midwest

American English - Northeastern

- American / New England
- American / Northeast
- American / Standard / Northeast
- American / Standard Northeast
- American Northeastern

The objection might be raised that these categorisations appear to be somewhat subjective, even though they are all employed by dialectologists and accent researchers (see, for instance, Trudgill & Hannah 2002: 45 on “Northern” and “Northern Cities”; Wells 1982: 518–527, Trudgill & Hannah 2002: 46 and Tottie 2002: 209 on the area variously referred to as “Northeast” or “eastern New England”; Gordon 2004b on the “Midwest”). After all, these labels may not necessarily refer to the same accents as respondents’ self-identifications. Nevertheless, it should be noted that fuzziness between categories has not been allowed to affect unduly the discussion, in 4.2 to 4.4, of any similarities between these accents and Dutch English. Since it is quite possible that some speakers who refer to themselves as “Midwestern” or “Northeastern” are in fact involved in the Northern Cities Chain Shift, whilst other speakers who have labelled themselves as “Northern” are not, the effects of this chain shift have been coded as affecting a minority of speakers from all three accent groups (see 4.4.2). As a result, the accent similarity coding for “Midwestern” and “Northern/Northern Cities” is identical, while “Northeastern” has been coded differently because of the characteristic feature of rhotacism in Eastern New England.

It is common practice for dialectologists and accent researchers to treat the accent of New York City separately, as in for example, Wells (1982: 401–418), Trudgill & Hannah (2002: 47) and Gordon (2004a: 284–289). In keeping with this, all respondents who identified themselves as being from “New York” were labelled as speakers of New York City English or “US/NY” (see Table 2.39). It is true that, strictly speaking, these informants could have been referring to anywhere in New York State, but references to areas outside the New York metropolitan area are often accompanied by the addition “upstate”. In fact, the single respondent (Subject 305) who used the term

“upstate NY” was consequently labelled “Northern/Northern Cities”, in keeping with Labov’s (1991: 14) description of the area involved in the Northern Cities Chain Shift. In addition, the social stigma attached to the New York City accent (see, for instance, Lippi-Green 1997: 175 and Wells 1982: 502) is such that anyone from New York State who does not have a New York City accent will probably wish to state this as unambiguously as possible.

Table 2.39. Accent self-identifications on the basis of which respondents were placed in the US/NY and US/WC minor accent groups.

<p>American English - New York City</p> <ul style="list-style-type: none"> • Amer / NY • New York • New York City / Long Island 	<p>American English - West and Southwest</p> <ul style="list-style-type: none"> • American (West Coast) • American / Southwest • American / Southwestern • American / Standard / West • American / Standard / West Coast • American / West Coast • American Standard (West Coast) • American West Coast • US West Coast • West Coast
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

All respondents using the term “West”, “West Coast”, “Southwest” or “Southwestern” were placed in the category “West & Southwest” or “US/WC” (see Table 2.39). Dialectologists tend not to distinguish sharply between Western and Southwestern American English. For instance, in C.K. Thomas’s classification of American speech areas (described in Wells 1982: 471–472), no provision is made for a separate West Coast accent; Californian English is simply included in the “Southwest”. More recently, Wolfram & Schilling-Estes (1998: 111) have also described the speech of Southern California as a “subdialect” of the Southwest. Non-linguists’ perceptions of dialect boundaries also show considerable overlap between Western and Southwestern accents, as has been demonstrated by Lance (1999: 303). Since the tokens in this experiment are unlikely to be pronounced differently in the West or the Southwest (however defined), these two groups of self-identifications have been subsumed under one label.

The other self-identifications provided by speakers from the United States are far less problematical. Any respondents referring to themselves as “Southern” were placed in the minor accent group of the same name, abbreviated to “US/SO” (see Table 2.40). A category such as “The south” is also used by, for instance, Wells (1982: 527–553) and Trudgill & Hannah (2002: 40–42). Similarly, any informants who refer to themselves as “East Coast” were placed in a minor accent group with the same name, coded as “US/EC”

(see Table 2.40). It was assumed that any respondents with distinctive Northeastern, New York City or Southeastern accents would not identify these with the rather generic label “East Coast” (which is not commonly used by linguists). As a result, the label “East Coast” was taken to exclude any speakers from the Northeast or the South, but to include speakers from the remaining areas on the Eastern seaboard. The unofficial term for this area is “Mid-Atlantic”, and normally includes New Jersey, Pennsylvania, Maryland, Washington DC and Delaware; it is largely coterminous with what some dialectologists define as the “Middle Atlantic area” (see, for instance, Wells 1982: 471–472, 487). This is why the single respondent (Subject 767) who referred to himself as “Mid-Atlantic” was also included under this heading.

Table 2.40. Accent self-identifications on the basis of which respondents were placed in the US/SO and US/EC minor accent groups.

American English - Southern	American English - East Coast
<ul style="list-style-type: none"> • American (Southern-influenced Standard) • American / Southern • American / Standard / St. Louis • American / Standard to Southern • American Southern • American Southern (Texas) • Amer-Southern • Southern 	<ul style="list-style-type: none"> • American / East Coast • American / fairly standard East Coast • American / Mid-Atlantic • American East Coast (Washington D.C.) • East Coast • East Coast North America

All Canadian respondents, irrespective of their provenance, were placed in a single Canadian category “CDN” (which serves as both a major and a minor accent group). This is because Canadian English, with the exception of Newfoundland, is “extremely homogeneous”, both in terms of “geographical [and] social variation” (Wells 1982: 491). As Brinton & Fee (2001: 423) put it, “the accents of Anglophone Canadians whose parents were born in Canada are nearly indistinguishable across the country”. The only important exception to this, the speech of Newfoundland, is not represented by a single informant, and therefore need not concern us here. The respondent (Subject 689) who described himself as “North American / West Coast” was also categorised as being Canadian, as the inclusive term “North American English” appears to be used much more commonly in Canada than in the US (see, for instance, McArthur 2002: 222).

Table 2.41. Accent self-identifications on the basis of which respondents were placed in the CDN major and minor accent group.

<p>Canadian English</p> <ul style="list-style-type: none"> • American / Canadian • Canada / South Central Ontario • Canadian • Canadian / American • Canadian / Quebec • Canadian / Standard • Canadian / West Coast 	<ul style="list-style-type: none"> • Canadian English • Canadian Standard • Canadian west coast • Canadian, Central • Central Canadian • North American / West Coast • Western Canadian
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

2.5.2 Analysis and categorisation of error assessments

The group of 545 judges produced a total number of 16,895 judgements (31 each, discounting the distractor and multiple submissions of single items). Prior to subjecting these to statistical analysis, they were analysed in order to establish in which cases an intended error had been detected clearly and unambiguously. There were three main ways in which respondents could detect an error successfully in the main body of the web survey. They could either click on the relevant segment or context, identify deviant “word stress” or “intonation”, or comment on any of these in the designated space (see 2.4.2). In those 10,628 cases (63%), when respondents had either detected the error successfully, or when it was assumed that they had done so, their original severity assessments had been retained. However, if participants had failed to identify an error clearly, their submissions for this had been recoded as “0”.

This section discusses the general principles applied in categorising submissions as either “unsuccessful” (recoded “0”) or “successful or potentially successful”. For instance, a segmental error was considered to have been detected successfully not only if a respondent had selected the relevant segment, but also if the error was mentioned in the space designated for comments (for a detailed discussion of participants’ comments, see 3.4.3). It should be noted that the position of some errors cannot easily be assigned to a single clickable segment (such as schwa-epenthesis in *FILM* or yod-dropping in *NEW*). In addition, certain other errors may also affect adjacent segments (such as incorrect vowel length in *ICE*, which may influence judges’ perception of final /s/; see Collins & Mees 2003b: 111). Some respondents appeared to be unduly distracted by the spelling of the context when selecting an error. For instance, 103 judges had located the error in *COLOUR* not in the single grapheme <o> in the initial syllable, but in the digraph <ou>. In view of this, it was decided that selecting any part of the relevant *context* of the error would also be considered a potential indication that this error had been detected successfully. (For an overview of these contexts, see Table 2.25 in 2.4.3.) Since most contexts were monosyllabic words (e.g. *bed*, *car*, *new* or *thin*), this resulted only in a few additional eligible segments, which were selected by only a small group of

judges.²⁷ If respondents mentioned specific contexts in their commentary, such “possibly relevant comments” (see 3.4.3) were also treated as successful submissions. If, however, participants discussed the appropriate error segment or context in their comments, but had simultaneously selected a different error in the clickable carrier sentence, their severity assessment was recoded to zero. In these 40 cases, it was clearly impossible to determine which of the errors they had evaluated. These constituted less than 1% of all judgements recoded to “0”.

A stress or suprasegmental error was also considered to have been identified successfully if respondents had either selected the appropriate category (“Word stress” or “Intonation”) or if they had made a statement to that effect in a separate comment (provided, of course, they had not identified an additional error at the same time). Although these error categories had been clearly defined in pop-up windows in the demo and the main body of the survey, there still appeared to be considerable confusion about the difference between them. For instance, no fewer than 26 judges had labelled the intended error in INT3 as a stress error, as opposed to 121 who correctly identified it as intonationally deviant. Similarly, while 14 respondents diagnosed an intonational problem in PERFECT, some of these also went on to describe it as a stress error in their comments. As a result, it was decided that any stress errors identified as intonationally deviant, as well as any intonation errors that were described as misplaced word stress, would be treated as successful submissions. In those cases where the stress or suprasegmental errors were located in the more restricted context of a word or phrase (as opposed to the entire carrier sentence, as in the case of intonation), selection of any of the appropriate segments was also treated as a successful submission.²⁸ The relevant contexts for these segments have been listed in Table 2.25 in 2.4.3. If any references to these segments or contexts were made in the space designated for comments, they were treated in the same way.

If respondents had identified a different error as opposed to, or in addition to, the intended one, either by selecting the error or by a statement in the comments, their severity assessment was recoded to the value of “0” (“not or incorrectly detected”). This was also done if they had replied “No” to the question “Does this sentence contain a clearly detectable error?” There were, however, two instances in which the severity coding “1”, originally assigned to this answer, was retained. Firstly, there is the somewhat ambiguous group of what may be referred to as the “clearly detectable non-errors”. There were

²⁷ For instance, if one considers the 275 respondents who had selected the error in BED as being located in the context *bed*, 266 of those had clicked on the grapheme <d>, as opposed to a mere eight for <e> and only one for . Similarly, in the considerably longer context *authority*, there were 496 submissions for the grapheme <th>, nine for <au>, three for the single grapheme <t>, and one for <o> and <r>.

²⁸ For instance, there were 144 respondents who identified the error in PERFECT as segmental. There were 74 submissions for the second grapheme <e>, 23 for the first <e>, 21 for <f>, 13 for <r>, 7 for <c>, 4 for <t> and 2 for <p>.

44 submissions in which judges mentioned one of the intended errors in their comments while simultaneously rating these as “No error”. It was assumed that this was a distinct category from those cases where an error had not been detected at all – or at least not demonstrably so. As a result, the severity value of 1 was retained (unless, of course, respondents had also identified an additional error). In this case, it may be seen as occupying an intermediate position between “0” (not or incorrectly detected) and “2” (“relatively unimportant”). Secondly, the severity value of 1 (“no error”) was also retained in the case of the distractor (INDIA). Since this carrier sentence did not contain any deviations from the RP or GA norm respectively, any other severity evaluations of this token were given the value of “0”. Such assessments occurred as a result of the tendency found in some judges to identify incorrectly an unintended “error”. It should be noted that assessments of the distractor have not been included in the analyses in Chapter 3 and 4, except where this provided insight into respondents’ behaviour (as in 3.5.15).

CHAPTER 3

THE NATIVE-SPEAKER EXPERIMENT: RESULTS AND ANALYSIS

3.1 Introduction and overall assessment of leniency and severity

3.1.1 Introduction

Once participants' responses in the online Native-speaker Experiment had been submitted, and their accent self-identifications and error severity judgements had been processed, these data were subjected to multi-level analysis using the MLwiN program (Rasbash *et al.* 2000; see 1.3.3). The results of this analysis are presented in this chapter, together with a detailed discussion of their implications.

The first sections (3.1.2 to 3.1.7) review any possible effects of the independent variables of age, sex, and accent group (both major and minor) on the dependent variables of respondents' self-identified leniency and their overall severity assessment of the tokens in the survey (excluding the distractor). In addition, this analysis encompasses any correlations between self-identified leniency and severity. While this section focuses on the effects on the overall severity assessment of all tokens, the subsequent section (3.2) reviews the overall severity assessment of the different tokens *in relation to each other*. This makes it possible to determine if tokens can be ranked by severity, as a result of which a hierarchy of error may be established. As is shown in 3.2.2 to 3.2.6, the results allow for the creation of three separate error hierarchies: one which is based on all severity judgements, and two additional hierarchies based on the two different forms of the experiment (the RP and GA versions respectively). The effect of independent variables of age and sex on the ranking of errors is discussed in section 3.3.

The focus in sections 3.4 and 3.5 is on a discussion of the severity estimates for each individual token. While sections 3.4.1 to 3.4.4 provide a general framework for this overview, a token-by-token analysis is provided in section 3.5. The general framework takes in: (1) a review of the effects of pairwise comparisons between major accent groups (3.4.1 and 3.4.2); (2) an overview of respondents' comments and their relation to the token-by-token analysis (3.4.3); and (3) a discussion of the effect of error detection rate on severity judgements (3.4.4). The third section shows how respondents' assessments of the severity of a particular error is affected by the question whether or not they had actually reported the error in the first place. As is shown in 3.4.4., it turns out that certain errors are assessed quite differently if only the judgements of respondents who

had in fact detected them are taken into consideration (the “adjusted severity”).¹ The results provided in sections 3.1 to 3.4 are subsequently brought to bear on individual discussions of the severity assessments of all 32 tokens (although, for the sake of convenience, a number of tokens are treated in the same subsections). This token-by-token analysis in 3.5 also compares and contrasts these results with severity assessments provided in a number of relevant textbooks and studies as discussed in 1.3.1. These include pronunciation manuals for Dutch learners of English such as Collins *et al.* (1987), Collins & Mees (1993, 2003b), and Gussenhoven & Broeders (1997) and a number of other relevant studies such as Brown (1988), Dretzke (1985), Jenkins (2000) and Koster & Koet (1993).

Section 3.6 discusses the severity assessments of individual errors in the Native-speaker Experiment as compared with those in the Dutch Experiment, insofar as this is justified by the similarities and differences between the two surveys. For the reasons discussed in 2.4.3 and in 3.6, this analysis only considers 22 of the 32 tokens, and only refers to respondents’ “adjusted severity”. The preliminary conclusions drawn from this comparison are also presented here. In the subsequent section (3.7), the results of all sections 3.1 to 3.6 are discussed in conjunction with each other, and reviewed in the light of the general aims, and more practical objectives, of this dissertation. A number of provisional conclusions are drawn from this, which are also set out in this section.

3.1.2 Self-identified leniency

At the end of the experiment, each respondent was asked to describe his or her attitude as “a judge of pronunciation” on a Likert scale ranging from 1 (“very lenient”) to 5 (“very strict”). As was pointed out in 2.4.2, these data were collected so as to normalise respondents’ assessments for their self-reported leniency, which varies between subjects. It also makes it possible to investigate how different groups of respondents perceived their own leniency and to establish any differences between these perceptions and their actual behaviour as judges. Multi-level analysis of these data showed that judges’ self-assessment of their leniency is not significantly affected by relevant variables such as sex, age, version of the experiment and major accent group (see 3.1.3). A more detailed analysis by minor accent group, however, revealed a trend for speakers from Scotland and the American East Coast to assess their own leniency differently from all the other groups combined (see 3.1.4).

¹ The term “adjusted severity” (AS) has been consistently used to refer to this specific variable. In all other cases, the general term “severity” is used to refer to the “composite” of detection and non-detection, i.e. with reference to *all* submissions for a particular token.

3.1.3 Self-identified leniency by sex, age, version and major accent group

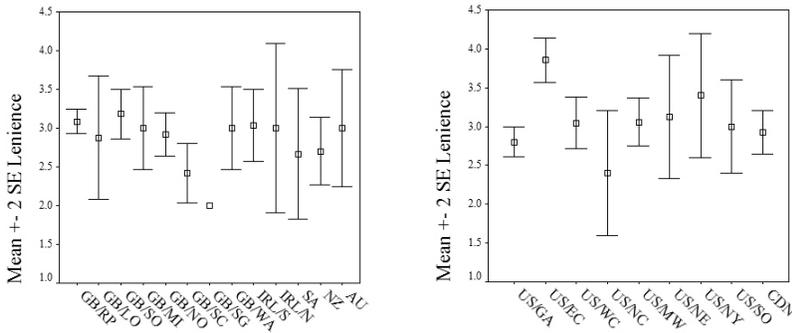
Analysis showed that the difference between the self-identified leniency for male respondents (mean 3.02; s.e. 0.1) and female respondents (mean 2.90; s.e. 0.1) is not significant ($\chi^2 < 1$, $df = 1$, n.s.). The regression coefficient for age (after centralisation) also turned out not to be significant (0.0008, s.e. 0.0031). The same was true for the difference between the self-identified leniency of native speakers opting for the RP version of the experiment vis-à-vis that of respondents taking the GA form ($\chi^2 < 1$, $df = 1$, n.s.). Multi-level modelling was also used to estimate self-identified leniency for a number of major accent groups. After the effects of sex and age for the respondents in each group had been subtracted, this resulted in the estimated means and standard errors listed in Table 3.1. Pairwise comparisons among accent groups showed that, after Bonferroni adjustment for multiple comparisons among $k = 7$ group means, none of the differences between these groups reached significance, neither at $\alpha = .05$ nor at a less strict $\alpha = .10$.

Table 3.1. Estimated means and standard errors for self-identified leniency, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	3.139	0.088
GB/NRP	2.988	0.091
IRL	3.103	0.219
AU&NZ&SA	2.797	0.173
US/GA	2.860	0.112
US/NGA	3.146	0.104
CDN	2.991	0.150

3.1.4 Self-identified leniency by minor accent group

If self-identified leniency is plotted by minor accent group, as in Figs. 3.1 and 3.2, there is considerable overlap between the error bars for the different groups. This indicates that there are no significant differences between them. However, two groups appear to have means that are strikingly different from all the other minor accent groups combined. These groups are (1) speakers of Scottish English (GB/SC); (2) judges who had identified themselves as being from the American East Coast (US/EC). As compared with judges from the rest of the English-speaking world, the GB/SC judges rate themselves as somewhat less strict, and the US/EC judges as slightly more severe. In a separate post-hoc test, the differences between the GB/SC judges and the rest of the English-speaking world turned out to be significant (Wald $Z = 2.3$, $p < .01$). Note that this was *not* the case with the US/EC respondents.



Figures 3.1 and 3.2. Means and error bars (2 standard errors) for self-identified leniency, broken down by minor accent group for RP (3.1) and GA (3.2) versions respectively.

3.1.5 Overall error severity assessment

When the severity scores for all tokens (always excluding the distractor) were subjected to multi-level analysis using the MLwiN program, it turned out that these were indeed affected by the variables sex, age and self-identified leniency, but not by version of the experiment or major accent group. More specifically, there appeared to be a tendency for older and female judges to be slightly more tolerant of the potential errors presented in the experiment. In addition, it became apparent that judges had, on the whole, identified their own leniency consistently and reliably (see 3.1.6). Furthermore, there was also a trend for judges from Australia, New Zealand, South Africa and Scotland to evaluate the potential errors less strictly than the other groups. The opposite was true for judges who had identified themselves as being from the American East Coast (see 3.1.7).

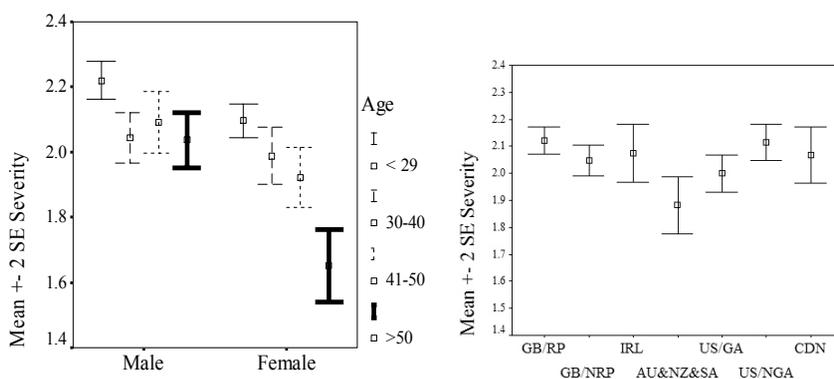
3.1.6 Overall severity assessment by sex, age, version and major accent group

Regression coefficients from the multi-level modelling in Table 3.2 show that the sex difference was significant at $\alpha = .05$ ($Z < -2$). The mean was estimated to be lower for female respondents by 0.134 scale points. The effect of age on severity is also significant ($Z < -2$, $p < .05$). An increase of one year in the age of a respondent leads to an estimated decrease in severity of 0.008. As has been shown, however, it is not only respondents' ageing but also their gender that affects their judgements. Figure 3.3 shows that it is in particular women over the age of 50 who are appreciably more tolerant of the potential errors produced by the actors in this experiment. In this context, it should perhaps be noted that the actors in both versions of the experiment were men (aged over 55). In addition, Table 3.2 demonstrates that a higher self-identified leniency score (implying an increase in self-diagnosed strictness) corresponds with an estimated increase of the mean severity by 0.163 ($Z = 7.21731$, $p < .05$). The positive correlation

between these variables shows that judges' assessment of their own leniency is in keeping with their actual strictness in evaluating the potential errors in this experiment.

Table 3.2. The effects of sex, age (centralised) and leniency on severity for all tokens excluding the distractor.

Predictor	Estimated effect on mean	Standard Error	Wald Z
SEX	-0.134	0.044	-3.1
AGE_C	-0.008	0.003	-5.0
LENIENCY	0.163	0.023	7.2



Figures 3.3 and 3.4. Means and error bars (2 standard errors) for severity, clustered and broken down by sex and age group (3.3), and broken down by major accent group (3.4) for all tokens excluding the distractor.

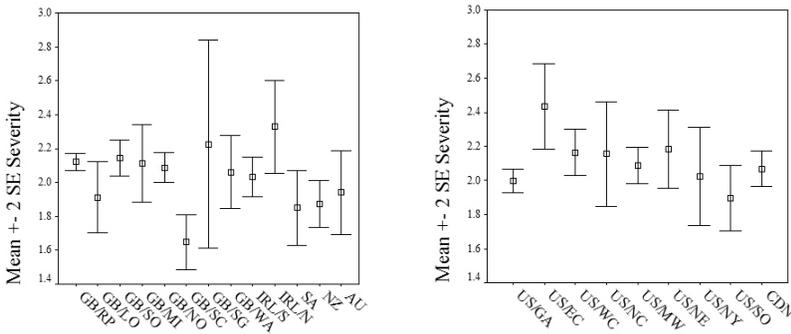
There was no significant difference in severity between those judges taking the RP version of the experiment as against those using the GA form ($\chi^2 < 1$, $df = 1$, n.s.). Similarly, pairwise comparisons among accent groups revealed that, after Bonferroni adjustment for multiple comparisons among $k = 7$ group means, none of the differences between these groups reached significance, neither at $\alpha = .05$ nor at a less strict $\alpha = .10$. Nor was there any significant variance among respondents within major accent groups. These results are plotted in Figure 3.4.

Interestingly, however, Figure 3.4 suggests that the Antipodeans (AU&NZ&SA) are slightly more tolerant of the potential errors in this experiment. This tendency was not significant in a multiple comparison among $k = 7$ groups, but a post-hoc two-way contrast between the Antipodeans and all the accent groups from the Northern hemisphere put together did yield a significant difference (Wald $Z = 3.8$, $p < .001$). When the scores for self-identified leniency were subjected to the same post-hoc two-way comparison, the difference was

also significant (Wald $Z = 3.6$, $p < .001$). In terms of severity for all tokens (again excluding the distractor), there are no significant differences between Australians (AU), New Zealanders (NZ) and South Africans (SA).

3.1.7 Overall error severity assessment by minor accent group

If the severity scores for the various minor accent groups are examined in more detail, it becomes clear that there are two other groups which appear to have judged the potential errors in this experiment differently from the rest. These are, firstly, the GB/SC speakers (see Figure 3.5) and, secondly, the US/EC group (see Figure 3.6). In a striking parallel with their self-identified leniency, the GB/SC judges turned out to be slightly less severe in their judgements, whereas the US/EC judges were slightly stricter. The post-hoc two-way contrast between the GB/SC speakers and those speaking other varieties of English was significant (Wald $Z = 2.5$, $p < .01$). It should be noted that these Scottish judges include neither the one self-identified bilingual speaker of Scottish English and Scots Gaelic nor any speakers of Ulster Scots, as these belong to demonstrably different accent groups. In actual fact, there is no statistical evidence to suggest that speakers of Northern and North-western Irish English (IRL/N) had judged the tokens any differently from the other groups. If the scores of the US/EC speakers are compared with those of all other groups in a similar post-hoc test, it similarly becomes apparent that these differences are significant (Wald $Z = 2.6$, $p < .01$).



Figures 3.5 and 3.6. Means and error bars (2 standard errors) for severity broken down by minor accent for all tokens except the distractor, for RP (3.5) and GA (3.6) versions respectively.

3.2 Overall assessment of individual tokens

3.2.1 Overall assessment

The severity scores for each of the 32 tokens have been subjected to multi-level analysis, resulting in one single overall severity estimate for each token, after subtracting the effects of respondents' age, sex, and leniency. The overall severity estimates for each of the 32 tokens are presented in Table 3.3 and are plotted in Figure 3.7.

It is immediately apparent from this diagram that the highest estimates are associated with errors involving stress placement and/or avoidance of weak forms (such as IMAGIN, PERFECT and TO_WALES) and, to a lesser extent, certain phonemic errors (including BED, BAT, VAN, WINE, THIN, AUTHOR and COLOUR) as well as a number of distributional or realisational errors (such as RED, DEAD and FILM). The lowest estimates, on the other hand, are connected with INT1, INT2 and INT3 (intonation), ICE, NEW and TELL (distributional or realisational differences) as well as the distractor INDIA. This would suggest that stress and stress-related phenomena appear to be judged most strictly, while the lowest estimates are associated with such diverse features as intonation, ICE, NEW and TELL.

Table 3.3. Overall severity estimates for all 32 tokens.

Token	Estimate	Standard Error
BED	3.057	0.056
BAT	2.958	0.061
VAN	3.117	0.060
WINE	2.745	0.061
THIN	3.416	0.050
AUTHOR	3.204	0.047
BOTH	2.315	0.072
OFF	1.115	0.060
THAT	1.196	0.066
WEATHER	2.480	0.067
BREATHE	2.280	0.080
RED	3.223	0.054
ICE	0.825	0.058
TIE	1.746	0.068
DEAD	2.878	0.066
FILM	2.993	0.060

CAR	1.792	0.063
HOT_TEA	1.174	0.070
INDIA	0.697	0.070
NEW	0.172	0.027
IMAGIN	3.646	0.040
PERFECT	3.759	0.036
TO_WALES	3.356	0.043
THAT_THA	1.292	0.066
SECONDARY	1.956	0.060
WOULD_ON	1.315	0.065
TELL	0.046	0.015
COLOUR	2.740	0.058
STOOD	2.317	0.068
INT1	1.97e-05	0.000
INT2	0.4746	0.045
INT3	0.7264	0.054

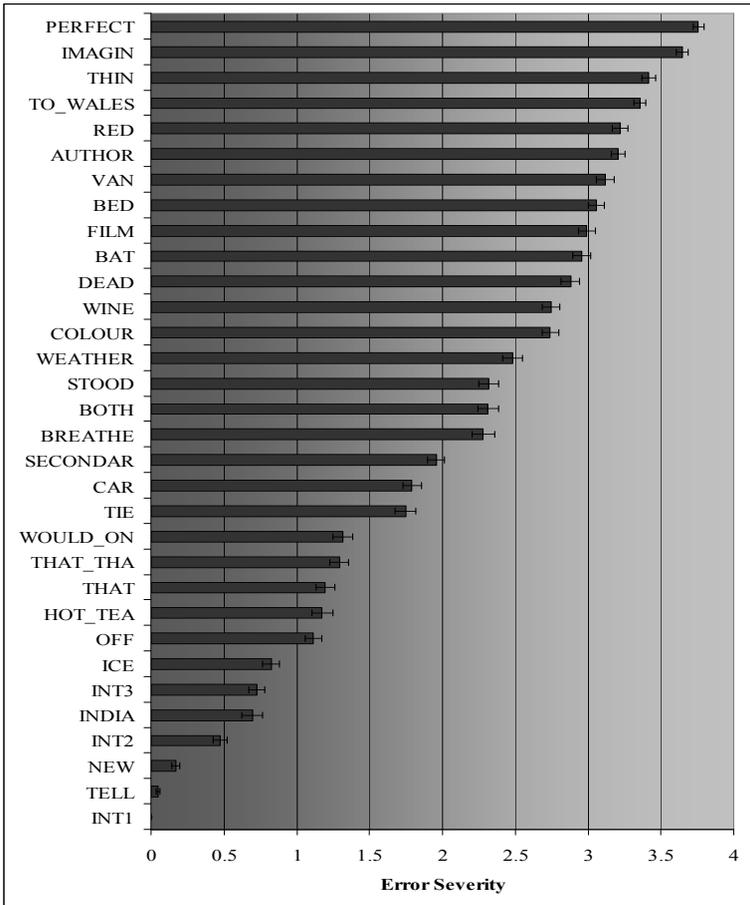


Figure 3.7. Bar chart with error bars showing overall error severity and standard errors for all 32 tokens.

3.2.2 Hierarchy of error

Before the estimates could be used to rank all 32 individual tokens in descending order of severity, so as to arrive at a “hierarchy of error”, Wald Z scores were calculated to establish if there are significant differences between the various estimates. Significance is obtained if $Z \geq |2|$, which was the case for most combinations of estimates (90%). Differences that turned out *not* to be significant, i.e. between two tokens that obtained similar severity scores, are highlighted in **bold** (see Tables 3.4 to 3.6).

Table 3.4. Wald Z scores for all possible combinations of severity estimates (tokens 1–11).

	1. BED	2. BAT	3. VAN	4. WINE	5. THIN	6. AUTHOR	7. BOTH	8. OFF	9. THAT	10. WEATHER	11. BREATHE
1. BED		0.8	-0.5	2.7	-3.4	-1.4	5.8	16.6	15.2	4.7	5.7
2. BAT	-0.8		-1.3	1.8	-4.1	-2.3	4.8	15.2	13.9	3.7	4.8
3. VAN	0.5	1.3		3.1	-2.7	-0.8	6.1	16.6	15.3	5.0	6.0
4. WINE	-2.7	-1.8	-3.1		-6.1	-4.2	3.2	13.5	12.2	2.1	3.3
5. THIN	3.4	4.1	2.7	6.1		2.2	9.0	20.9	19.2	8.1	8.8
6. AUTHOR	1.4	2.3	0.8	4.2	-2.2		7.4	19.4	17.7	6.4	7.3
7. BOTH	-5.8	-4.8	-6.1	-3.2	-9.0	-7.4		9.0	8.1	-1.2	0.2
8. OFF	-16.6	-15.2	-16.6	-13.5	-20.9	-19.4	-9.0		-0.6	-10.8	-8.3
9. THAT	-15.2	-13.9	-15.3	-12.2	-19.2	-17.7	-8.1	0.6		-9.7	-7.5
10. WEATHER	-4.7	-3.7	-5.0	-2.1	-8.1	-6.4	1.2	10.8	9.7		1.4
11. BREATHE	-5.7	-4.8	-6.0	-3.3	-8.8	-7.3	-0.2	8.3	7.5	-1.4	
12. RED	1.5	2.3	0.9	4.2	-1.9	0.2	7.2	18.4	16.9	6.1	7.0
13. ICE	-19.6	-18.0	-19.5	-16.2	-24.2	-22.6	-11.5	-2.5	-3.0	-13.3	-10.6
14. TIE	-10.5	-9.4	-10.7	-7.7	-14.1	-12.6	-4.0	4.9	4.1	-5.4	-3.6
15. DEAD	-1.5	-0.6	-1.9	1.0	-4.6	-2.9	4.1	13.9	12.7	3.0	4.1
16. FILM	-0.6	0.3	-1.0	2.1	-3.9	-2.0	5.1	15.6	14.3	4.1	5.1
17. CAR	-10.6	-9.4	-10.8	-7.7	-14.4	-12.8	-3.9	5.5	4.6	-5.3	-3.4
18. HOT TEA	-14.9	-13.6	-14.9	-12.0	-18.7	-17.2	-8.0	0.5	-0.2	-9.5	-7.4
19. INDIA	-18.6	-17.2	-18.6	-15.6	-22.7	-21.3	-11.3	-3.2	-3.7	-13.0	-10.6
20. NEW	-34.5	-31.6	-33.7	-29.2	-42.2	-40.6	-21.5	-10.8	-11.0	-24.6	-19.7
21. IMAGIN	6.1	6.8	5.3	9.0	2.6	5.1	11.9	25.3	23.2	11.0	11.5
22. PERFECT	7.6	8.3	6.7	10.5	4.0	6.6	13.3	27.4	25.1	12.5	12.8
23. TO WALES	3.0	3.8	2.3	5.9	-0.7	1.7	9.1	21.8	19.9	8.0	8.8
24. THAT THA	-14.4	-13.1	-14.4	-11.4	-18.3	-16.8	-7.4	1.4	0.7	-8.9	-6.8
25. SECONDAR	-9.5	-8.3	-9.7	-6.5	-13.3	-11.6	-2.7	7.0	6.0	-4.1	-2.3
26. WOULD ON	-14.3	-13.0	-14.4	-11.4	-18.3	-16.8	-7.3	1.6	0.9	-8.8	-6.7
27. TELL	-42.2	-38.4	-41.0	-35.6	-52.2	-50.6	-26.0	-14.2	-14.2	-29.8	-23.6
28. COLOUR	-2.8	-1.8	-3.2	0.0	-6.3	-4.4	3.3	13.7	12.5	2.1	3.3
29. STOOD	-5.9	-5.0	-6.2	-3.3	-9.3	-7.7	0.0	9.4	8.4	-1.2	0.3
30. INT1	-54.0	-48.3	-51.7	-44.9	-68.5	-67.1	-31.9	-18.4	-18.1	-37.1	-28.5
31. INT2	-25.4	-23.4	-25.1	-21.4	-31.0	-29.5	-15.7	-6.1	-6.5	-17.9	-14.5
32. INT3	-21.2	-19.5	-21.0	-17.6	-26.0	-24.5	-12.6	-3.4	-3.9	-14.6	-11.7

Table 3.5. Wald Z scores for all possible combinations of severity estimates (tokens 12–22).

	12. RED	13. ICE	14. TIE	15. DEAD	16. FILM	17. CAR	18. HOT_TEA	19. INDIA	20. NEW	21. IMAGIN	22. PERFECT
1. BED	-1.5	19.6	10.5	1.5	0.6	10.6	14.9	18.6	34.5	-6.1	-7.6
2. BAT	-2.3	18.0	9.4	0.6	-0.3	9.4	13.6	17.2	31.6	-6.8	-8.3
3. VAN	-0.9	19.5	10.7	1.9	1.0	10.8	14.9	18.6	33.7	-5.3	-6.7
4. WINE	-4.2	16.2	7.7	-1.0	-2.1	7.7	12.0	15.6	29.2	-9.0	-10.5
5. THIN	1.9	24.2	14.1	4.6	3.9	14.4	18.7	22.7	42.2	-2.6	-4.0
6. AUTHOR	-0.2	22.6	12.6	2.9	2.0	12.8	17.2	21.3	40.6	-5.1	-6.6
7. BOTH	-7.2	11.5	4.0	-4.1	-5.1	3.9	8.0	11.3	21.5	-11.9	-13.3
8. OFF	-18.4	2.5	-4.9	-13.9	-15.6	-5.5	-0.5	3.2	10.8	-25.3	-27.4
9. THAT	-16.9	3.0	-4.1	-12.7	-14.3	-4.6	0.2	3.7	11.0	-23.2	-25.1
10. WEATHER	-6.1	13.3	5.4	-3.0	-4.1	5.3	9.5	13.0	24.6	-11.0	-12.5
11. BREATHE	-7.0	10.6	3.6	-4.1	-5.1	3.4	7.4	10.6	19.7	-11.5	-12.8
12. RED		21.4	12.0	2.9	2.0	12.2	16.4	20.3	37.4	-4.5	-5.9
13. ICE	-21.4		-7.3	-16.5	-18.5	-8.0	-2.7	1.0	7.7	-29.0	-31.3
14. TIE	-12.0	7.3		-8.4	-9.7	-0.3	4.1	7.6	16.4	-17.6	-19.3
15. DEAD	-2.9	16.5	8.4		-0.9	8.4	12.5	16.0	28.9	-7.2	-8.6
16. FILM	-2.0	18.5	9.7	0.9		9.8	14.0	17.7	32.4	-6.6	-8.0
17. CAR	-12.2	8.0	0.3	-8.4	-9.8		4.6	8.2	17.9	-18.0	-19.8
18. HOT TEA	-16.4	2.7	-4.1	-12.5	-14.0	-4.6		3.4	10.3	-22.5	-24.3
19. INDIA	-20.3	-1.0	-7.6	-16.0	-17.7	-8.2	-3.4		5.4	-26.8	-28.8
20. NEW	-37.4	-7.7	-16.4	-28.9	-32.4	-17.9	-10.3	-5.4		-52.0	-56.6
21. IMAGIN	4.5	29.0	17.6	7.2	6.6	18.0	22.5	26.8	52.0		-1.5
22. PERFECT	5.9	31.3	19.3	8.6	8.0	19.8	24.3	28.8	56.6	1.5	
23. TO WALES	1.4	25.2	14.5	4.4	3.5	14.8	19.3	23.5	45.5	-3.5	-5.1
24. THAT THA	-16.0	3.8	-3.4	-11.9	-13.5	-3.9	0.9	4.3	11.9	-22.2	-24.1
25. SECONDAR	-11.1	9.6	1.6	-7.3	-8.7	1.3	6.0	9.7	20.4	-17.0	-18.7
26. WOULD ON	-16.0	4.0	-3.2	-11.9	-13.4	-3.7	1.0	4.6	12.4	-22.2	-24.1
27. TELL	-45.8	-10.7	-20.4	-34.8	-39.5	-22.3	-13.2	-7.6	-3.0	-66.0	-72.7
28. COLOUR	-4.3	16.5	7.9	-1.1	-2.1	7.8	12.2	15.9	30.0	-9.3	-10.8
29. STOOD	-7.4	11.9	4.2	-4.2	-5.3	4.0	8.3	11.7	22.5	-12.3	-13.8
30. INT1	-59.0	-14.2	-25.4	-43.2	-49.9	-28.2	-16.6	-9.9	-6.2	-91.4	-103.3
31. INT2	-27.6	-3.4	-11.2	-21.5	-24.0	-12.1	-6.1	-1.9	4.2	-37.4	-40.4
32. INT3	-23.1	-0.9	-8.3	-17.9	-20.0	-9.1	-3.6	0.2	6.8	-31.3	-33.8

Table 3.6. Wald Z scores for all possible combinations of severity estimates (tokens 23–32).

	23. TO_WALES	24. THAT_THA	25. SECONDAR	26. WOULD_ON	27. TELL	28. COLOUR	29. STOOD	30. INT1	31. INT2	32. INT3
1. BED	-3.0	14.4	9.5	14.3	42.2	2.8	5.9	54.0	25.4	21.2
2. BAT	-3.8	13.1	8.3	13.0	38.4	1.8	5.0	48.3	23.4	19.5
3. VAN	-2.3	14.4	9.7	14.4	41.0	3.2	6.2	51.7	25.1	21.0
4. WINE	-5.9	11.4	6.5	11.4	35.6	0.0	3.3	44.9	21.4	17.6
5. THIN	0.7	18.3	13.3	18.3	52.2	6.3	9.3	68.5	31.0	26.0
6. AUTHOR	-1.7	16.8	11.6	16.8	50.6	4.4	7.7	67.1	29.5	24.5
7. BOTH	-9.1	7.4	2.7	7.3	26.0	-3.3	0.0	31.9	15.7	12.6
8. OFF	-21.8	-1.4	-7.0	-1.6	14.2	-13.7	-9.4	18.4	6.1	3.4
9. THAT	-19.9	-0.7	-6.0	-0.9	14.2	-12.5	-8.4	18.1	6.5	3.9
10. WEATHER	-8.0	8.9	4.1	8.8	29.8	-2.1	1.2	37.1	17.9	14.6
11. BREATHE	-8.8	6.8	2.3	6.7	23.6	-3.3	-0.3	28.5	14.5	11.7
12. RED	-1.4	16.0	11.1	16.0	45.8	4.3	7.4	59.0	27.6	23.1
13. ICE	-25.2	-3.8	-9.6	-4.0	10.7	-16.5	-11.9	14.2	3.4	0.9
14. TIE	-14.5	3.4	-1.6	3.2	20.4	-7.9	-4.2	25.4	11.2	8.3
15. DEAD	-4.4	11.9	7.3	11.9	34.8	1.1	4.2	43.2	21.5	17.9
16. FILM	-3.5	13.5	8.7	13.4	39.5	2.1	5.3	49.9	24.0	20.0
17. CAR	-14.8	3.9	-1.3	3.7	22.3	-7.8	-4.0	28.2	12.1	9.1
18. HOT TEA	-19.3	-0.9	-6.0	-1.0	13.2	-12.2	-8.3	16.6	6.1	3.6
19. INDIA	-23.5	-4.3	-9.7	-4.6	7.6	-15.9	-11.7	9.9	1.9	-0.2
20. NEW	-45.5	-11.9	-20.4	-12.4	3.0	-30.0	-22.5	6.2	-4.2	-6.8
21. IMAGIN	3.5	22.2	17.0	22.2	66.0	9.3	12.3	91.4	37.4	31.3
22. PERFECT	5.1	24.1	18.7	24.1	72.7	10.8	13.8	103.3	40.4	33.8
23. TO WALES		18.9	13.6	18.9	57.5	6.1	9.4	78.1	32.8	27.3
24. THAT_THA	-18.9		-5.2	-0.2	15.3	-11.6	-7.6	19.4	7.3	4.7
25. SECONDAR	-13.6	5.2		5.1	25.4	-6.6	-2.8	32.4	14.1	10.8
26. WOULD ON	-18.9	0.2	-5.1		15.8	-11.5	-7.5	20.1	7.6	4.9
27. TELL	-57.5	-15.3	-25.4	-15.8		-36.8	-27.3	3.0	-7.1	-9.9
28. COLOUR	-6.1	11.6	6.6	11.5	36.8		3.3	46.8	21.9	18.0
29. STOOD	-9.4	7.6	2.8	7.5	27.3	-3.3		33.9	16.3	13.1
30. INT1	-78.1	-19.4	-32.4	-20.1	-3.0	-46.8	-33.9		-10.4	-13.4
31. INT2	-32.8	-7.3	-14.1	-7.6	7.1	-21.9	-16.3	10.4		-2.5
32. INT3	-27.3	-4.7	-10.8	-4.9	9.9	-18.0	-13.1	13.4	2.5	

Tables 3.4 to 3.6 show that, while most estimates differ significantly from one another, a number are clustered together so closely that the differences between them are negligible. In Table 3.4, Column 2, for instance, the estimates for BAT, VAN, AUTHOR, RED, DEAD, FILM have been highlighted in **bold**, which means they are too similar to BED to be significantly different. Similarly, VAN does not differ significantly from BED, BAT, AUTHOR, DEAD and FILM. This would suggest that BED and VAN should be ranked equally high, even though the severity estimate for VAN is 3.117, as opposed to 3.057 for BED. Even if it is not entirely possible to rank all individual 32 tokens in descending order of severity, so as to arrive at a true “hierarchy of error”, tokens with similar values can be ranked into 9 discrete clusters. The differences between adjacent tokens in a cluster are not significant. These results are plotted in Table 3.7.

Table 3.7. Ranking of tokens by clusters of adjacent severity scores.

Token	Estimate	Error Category
(1) PERFECT	3.759	Stress
(1) IMAGIN	3.646	Stress
(2) THIN	3.416	Phonemic
(2) TO_WALES	3.356	Suprasegmental
(2) RED	3.223	Realisational
(2) AUTHOR	3.204	Phonemic
(2) VAN	3.117	Phonemic
(2) BED	3.057	Phonemic
(2) FILM	2.993	Distributional
(2) BAT	2.958	Phonemic
(2) DEAD	2.878	Realisational
(2) WINE	2.745	Phonemic
(2) COLOUR	2.740	Phonemic
(3) WEATHER	2.480	Phonemic
(3) STOOD	2.317	Phonemic
(3) BOTH	2.315	Phonemic
(3) BREATHE	2.280	Phonemic
(4) SECONDAR	1.956	Suprasegmental
(4) CAR	1.792	Distributional
(4) TIE	1.746	Realisational
(5) WOULD_ON	1.315	Suprasegmental

(5) THAT_THA	1.292	Suprasegmental
(5) THAT	1.196	Phonemic
(5) HOT_TEA	1.174	Distributional
(5) OFF	1.115	Phonemic
(6) ICE	0.825	Realisational
(6) INT3	0.726	Suprasegmental
(6) INDIA	0.697	Distractor
(6) INT2	0.475	Suprasegmental
(7) NEW	0.172	Distributional
(8) TELL	0.046	Realisational
(9) INT1	0.000	Suprasegmental

Table 3.7 clearly reveals a hierarchy of error for 9 discrete clusters. While the highest estimates (> 3.5) are associated with stress placement, the lowest estimates (< 1.0) involve intonation and certain distributional/realisational errors (which Johansson 1975 would refer to as “sub-phonemic”). In between these extremes, the “upper intermediate” ranges (3.5–2.0) consist of virtually all the potential phonemic errors, two highly stigmatised distributional and realisational differences (RED and FILM) and two pronunciations errors that, if classified on the basis of their potential effect on native speakers, could also be considered phonemic (DEAD) or stress-related errors (TO_WALES), while the “lower intermediate” ranges (2.0–1.0) consist of both suprasegmental and “sub-phonemic” errors as well as two potential phonemic errors that occur in high-frequency grammar words (BOTH and THAT). These results will be discussed in more detail in the token-by-token analysis (3.5).

3.2.3 Overall assessment of individual tokens by version of the experiment

Even though there were no significant differences in overall severity assessment between the RP and GA versions of the experiment, a multi-level analysis of overall assessment of *individual* tokens reveals that the majority of these had been judged demonstrably differently in the two versions. This makes it possible to adjust the overall hierarchy of error for these two major versions of Standard English.

As is described in 3.2.4 and 3.2.5, the different severity estimates for each token were distributed quite similarly in the RP and GA versions, with the highest means associated with stress and with certain phonemic differences, and the lowest means with intonation, TELL (in the case of RP) and NEW (in the case of GA). Since there were fewer significant differences between estimates in either the RP or the GA version, it was more difficult to rank these individually

for each version with any degree of confidence. In addition, a further comparison of the RP and GA estimates by token shows that only 10 tokens had been assessed very similarly, whereas the estimates for as many as 22 tokens were statistically significantly different in the two versions (see 3.2.6). While the severity estimates for BAT, WINE, AUTHOR, BREATHE, FILM and PERFECT were consistently high for both versions, consistent but lower scores were associated with HOT_TEA, SECONDAR, as well as with two intonation tokens. There was significant inter-version variation for the other tokens, notably those that differed by more than one Likert scale point. Some of these included tokens featuring either /θ/ or /ð/.

3.2.4 Overall assessment of individual tokens in the RP version

For the RP version, severity estimates and standard errors were calculated for each token using the MLwiN program. These are presented in Table 3.8, and are also plotted in Figure 3.8 below. Similarly to the overall severity estimates, the highest means are associated with stress placement and/or avoidance of weak forms as well as certain phonemic contrasts and distributional or realisational differences. In the RP version, the lowest means are also connected with intonation and TELL (but not NEW).

Table 3.8. Overall severity estimates for all 32 tokens (RP version only).

Token	Estimate	Standard Error
BED	3.018	0.089
BAT	3.004	0.093
VAN	2.938	0.097
WINE	2.713	0.093
THIN	3.468	0.070
AUTHOR	3.250	0.070
BOTH	1.582	0.108
OFF	0.981	0.093
THAT	0.929	0.097
WEATHER	1.642	0.101
BREATHE	2.265	0.124
RED	3.093	0.083
ICE	1.252	0.092
TIE	2.300	0.098
DEAD	2.603	0.108
FILM	2.946	0.094

CAR	1.736	0.097
HOT_TEA	1.400	0.108
INDIA	0.620	0.032
NEW	1.457	0.082
IMAGIN	3.592	0.060
PERFECT	3.778	0.058
TO_WALES	3.152	0.072
THAT_THA	1.517	0.105
SECONDAR	2.002	0.094
WOULD_ON	2.030	0.100
TELL	3.13e-17	6.86e-10
COLOUR	3.125	0.073
STOOD	2.918	0.087
INT1	0.166	0.044
INT2	0.402	0.068
INT3	0.657	0.085

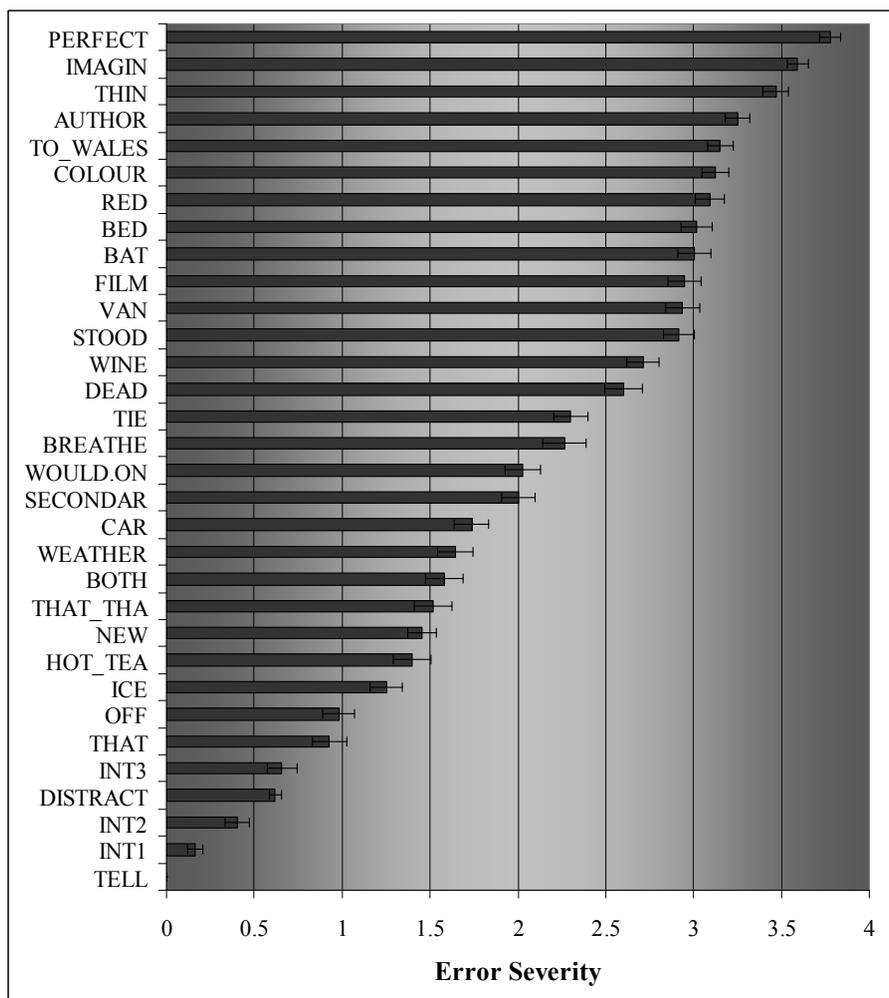


Figure 3.8. Bar chart with error bars showing overall severity and standard errors for all 32 tokens (RP version).

For the RP version of the experiment, Wald Z scores were calculated to determine if there are significant differences between the various estimates. Significance ($Z \geq |2|$) was obtained in relatively fewer cases than with the overall severity estimates (82% as opposed to 90%). This makes any attempts to rank these tokens in descending order of severity a little less reliable. Instances where there is *no* significant difference between any two estimates are highlighted in **bold** (see Tables 3.9, 3.10 and 3.11).

Table 3.9. Wald Z scores for all possible combinations of severity estimates in the RP version of the experiment (tokens 1–11).

	1. BED	2. BAT	3. VAN	4. WINE	5. THIN	6. AUTHOR	7. BOTH	8. OFF	9. THAT	10. WEATHER	11. BREATHE
1. BED		0.1	0.4	1.7	-2.8	-1.5	7.3	11.2	11.2	7.2	3.5
2. BAT	-0.1		0.3	1.6	-2.8	-1.5	7.1	10.9	10.9	7.0	3.4
3. VAN	-0.4	-0.3		1.2	-3.2	-1.9	6.6	10.3	10.3	6.5	3.0
4. WINE	-1.7	-1.6	-1.2		-4.6	-3.3	5.6	9.3	9.4	5.5	2.1
5. THIN	2.8	2.8	3.2	4.6		1.5	10.6	15.2	15.1	10.6	6.2
6. AUTHOR	1.5	1.5	1.9	3.3	-1.5		9.4	13.9	13.8	9.4	5.1
7. BOTH	-7.3	-7.1	-6.6	-5.6	-10.6	-9.4		3.0	3.2	-0.3	-3.0
8. OFF	-11.2	-10.9	-10.3	-9.3	-15.2	-13.9	-3.0		0.3	-3.4	-5.9
9. THAT	-11.2	-10.9	-10.3	-9.4	-15.1	-13.8	-3.2	-0.3		-3.6	-6.0
10. WEATHER	-7.2	-7.0	-6.5	-5.5	-10.6	-9.4	0.3	3.4	3.6		-2.8
11. BREATHE	-3.5	-3.4	-3.0	-2.1	-6.2	-5.1	3.0	5.9	6.0	2.8	
12. RED	0.4	0.5	0.9	2.2	-2.4	-1.0	7.9	12.0	12.0	7.9	4.0
13. ICE	-9.8	-9.5	-8.9	-7.9	-13.7	-12.3	-1.7	1.5	1.7	-2.0	-4.7
14. TIE	-3.8	-3.7	-3.3	-2.2	-6.9	-5.6	3.5	6.9	7.0	3.3	0.2
15. DEAD	-2.1	-2.0	-1.6	-0.5	-4.9	-3.6	4.7	8.1	8.2	4.6	1.5
16. FILM	-0.4	-0.3	0.0	1.2	-3.2	-1.8	6.7	10.5	10.5	6.7	3.1
17. CAR	-6.9	-6.7	-6.2	-5.1	-10.4	-9.0	0.8	4.0	4.2	0.5	-2.4
18. HOT TEA	-8.2	-8.0	-7.5	-6.5	-11.6	-10.3	-0.8	2.1	2.3	-1.2	-3.7
19. INDIA	-19.9	-19.1	-17.9	-16.7	-27.7	-25.6	-6.9	-2.9	-2.4	-7.7	-10.5
20. NEW	-9.1	-8.8	-8.3	-7.2	-13.2	-11.8	-0.7	2.7	2.9	-1.0	-3.9
21. IMAGIN	3.9	3.9	4.2	5.7	1.0	2.6	12.0	17.1	17.0	12.1	7.2
22. PERFECT	5.2	5.1	5.4	7.0	2.4	4.1	13.2	18.5	18.3	13.4	8.3
23. TO WALES	0.8	0.9	1.3	2.7	-2.2	-0.7	8.8	13.2	13.2	8.7	4.5
24. THAT THA	-7.8	-7.5	-7.0	-6.0	-11.1	-9.9	-0.3	2.7	2.9	-0.6	-3.3
25. SECONDAR	-5.6	-5.4	-4.9	-3.8	-8.9	-7.6	2.1	5.5	5.6	1.8	-1.2
26. WOULD ON	-5.2	-5.1	-4.6	-3.5	-8.5	-7.2	2.2	5.5	5.6	1.9	-1.1
27. TELL	-34.1	-32.4	-30.3	-29.1	-49.2	-46.1	-14.7	-10.6	-9.5	-16.2	-18.3
28. COLOUR	0.7	0.7	1.1	2.5	-2.4	-0.9	8.5	12.9	12.9	8.5	4.4
29. STOOD	-0.6	-0.5	-0.1	1.1	-3.5	-2.1	6.8	10.7	10.8	6.8	3.1
30. INT1	-21.5	-20.7	-19.6	-18.5	-28.7	-26.9	-9.3	-5.9	-5.4	-10.1	-12.5
31. INT2	-16.7	-16.2	-15.4	-14.3	-22.2	-20.6	-6.7	-3.6	-3.2	-7.3	-9.7
32. INT3	-13.6	-13.2	-12.5	-11.5	-18.1	-16.7	-4.8	-1.8	-1.5	-5.3	-7.7

Table 3.10. Wald Z scores for all possible combinations of severity estimates in the RP version of the experiment (tokens 12–22).

	12. RED	13. ICE	14. TIE	15. DEAD	16. FILM	17. CAR	18. HOT_TEA	19. INDIA	20. NEW	21. IMAGIN	22. PERFECT
1. BED	-0.4	9.8	3.8	2.1	0.4	6.9	8.2	19.9	9.1	-3.9	-5.2
2. BAT	-0.5	9.5	3.7	2.0	0.3	6.7	8.0	19.1	8.8	-3.9	-5.1
3. VAN	-0.9	8.9	3.3	1.6	0.0	6.2	7.5	17.9	8.3	-4.2	-5.4
4. WINE	-2.2	7.9	2.2	0.5	-1.2	5.1	6.5	16.7	7.2	-5.7	-7.0
5. THIN	2.4	13.7	6.9	4.9	3.2	10.4	11.6	27.7	13.2	-1.0	-2.4
6. AUTHOR	1.0	12.3	5.6	3.6	1.8	9.0	10.3	25.6	11.8	-2.6	-4.1
7. BOTH	-7.9	1.7	-3.5	-4.7	-6.7	-0.8	0.8	6.9	0.7	-12.0	-13.2
8. OFF	-12.0	-1.5	-6.9	-8.1	-10.5	-4.0	-2.1	2.9	-2.7	-17.1	-18.5
9. THAT	-12.0	-1.7	-7.0	-8.2	-10.5	-4.2	-2.3	2.4	-2.9	-17.0	-18.3
10. WEATHER	-7.9	2.0	-3.3	-4.6	-6.7	-0.5	1.2	7.7	1.0	-12.1	-13.4
11. BREATHE	-4.0	4.7	-0.2	-1.5	-3.1	2.4	3.7	10.5	3.9	-7.2	-8.3
12. RED		10.5	4.4	2.6	0.8	7.6	8.9	21.5	9.9	-3.5	-4.9
13. ICE	-10.5		-5.5	-6.8	-9.1	-2.6	-0.7	5.1	-1.2	-15.5	-16.8
14. TIE	-4.4	5.5		-1.5	-3.4	2.9	4.4	12.9	4.7	-8.2	-9.4
15. DEAD	-2.6	6.8	1.5		-1.7	4.2	5.6	14.2	6.0	-5.9	-7.1
16. FILM	-0.8	9.1	3.4	1.7		6.3	7.6	18.4	8.4	-4.2	-5.4
17. CAR	-7.6	2.6	-2.9	-4.2	-6.3		1.6	8.7	1.6	-11.9	-13.2
18. HOT TEA	-8.9	0.7	-4.4	-5.6	-7.6	-1.6		5.5	-0.3	-13.0	-14.3
19. INDIA	-21.5	-5.1	-12.9	-14.2	-18.4	-8.7	-5.5		-7.3	-32.4	-34.9
20. NEW	-9.9	1.2	-4.7	-6.0	-8.4	-1.6	0.3	7.3		-15.1	-16.5
21. IMAGIN	3.5	15.5	8.2	5.9	4.2	11.9	13.0	32.4	15.1		-1.6
22. PERFECT	4.9	16.8	9.4	7.1	5.4	13.2	14.3	34.9	16.5	1.6	
23. TO WALES	0.4	11.6	5.0	3.1	1.2	8.4	9.7	24.4	11.0	-3.4	-4.8
24. THAT THA	-8.4	1.3	-3.9	-5.1	-7.2	-1.1	0.5	6.5	0.3	-12.6	-13.8
25. SECONDAR	-6.2	4.0	-1.6	-3.0	-5.0	1.4	3.0	11.0	3.1	-10.4	-11.7
26. WOULD ON	-5.8	4.1	-1.4	-2.8	-4.7	1.5	3.0	10.7	3.2	-9.8	-11.1
27. TELL	-37.3	-13.7	-23.4	-24.1	-31.2	-17.9	-12.9	-19.3	-17.7	-60.2	-64.8
28. COLOUR	0.2	11.3	4.8	2.9	1.1	8.2	9.5	23.7	10.7	-3.5	-5.0
29. STOOD	-1.0	9.3	3.3	1.6	-0.2	6.4	7.8	19.2	8.6	-4.6	-5.9
30. INT1	-23.0	-8.0	-15.0	-16.0	-20.0	-11.1	-8.1	-5.9	-10.2	-32.9	-35.2
31. INT2	-17.9	-5.3	-11.4	-12.5	-15.7	-8.1	-5.7	-2.2	-7.0	-25.0	-26.7
32. INT3	-14.5	-3.4	-9.0	-10.1	-12.8	-5.9	-3.8	0.3	-4.8	-20.3	-21.8

Table 3.11. Wald Z scores for all possible combinations of severity estimates in the RP version of the experiment (tokens 23–32).

	23. TO_WALES	24. THAT_THA	25. SECONDAR	26. WOULD_ON	27. TELL	28. COLOUR	29. STOOD	30. INT1	31. INT2	32. INT3
1. BED	-0.8	7.8	5.6	5.2	34.1	-0.7	0.6	21.5	16.7	13.6
2. BAT	-0.9	7.5	5.4	5.1	32.4	-0.7	0.5	20.7	16.2	13.2
3. VAN	-1.3	7.0	4.9	4.6	30.3	-1.1	0.1	19.6	15.4	12.5
4. WINE	-2.7	6.0	3.8	3.5	29.1	-2.5	-1.1	18.5	14.3	11.5
5. THIN	2.2	11.1	8.9	8.5	49.2	2.4	3.5	28.7	22.2	18.1
6. AUTHOR	0.7	9.9	7.6	7.2	46.1	0.9	2.1	26.9	20.6	16.7
7. BOTH	-8.8	0.3	-2.1	-2.2	14.7	-8.5	-6.8	9.3	6.7	4.8
8. OFF	-13.2	-2.7	-5.5	-5.5	10.6	-12.9	-10.7	5.9	3.6	1.8
9. THAT	-13.2	-2.9	-5.6	-5.6	9.5	-12.9	-10.8	5.4	3.2	1.5
10. WEATHER	-8.7	0.6	-1.8	-1.9	16.2	-8.5	-6.8	10.1	7.3	5.3
11. BREATHE	-4.5	3.3	1.2	1.1	18.3	-4.4	-3.1	12.5	9.7	7.7
12. RED	-0.4	8.4	6.2	5.8	37.3	-0.2	1.0	23.0	17.9	14.5
13. ICE	-11.6	-1.3	-4.0	-4.1	13.7	-11.3	-9.3	8.0	5.3	3.4
14. TIE	-5.0	3.9	1.6	1.4	23.4	-4.8	-3.3	15.0	11.4	9.0
15. DEAD	-3.1	5.1	3.0	2.8	24.1	-2.9	-1.6	16.0	12.5	10.1
16. FILM	-1.2	7.2	5.0	4.7	31.2	-1.1	0.2	20.0	15.7	12.8
17. CAR	-8.4	1.1	-1.4	-1.5	17.9	-8.2	-6.4	11.1	8.1	5.9
18. HOT TEA	-9.7	-0.5	-3.0	-3.0	12.9	-9.5	-7.8	8.1	5.7	3.8
19. INDIA	-24.4	-6.5	-11.0	-10.7	19.3	-23.7	-19.2	5.9	2.2	-0.3
20. NEW	-11.0	-0.3	-3.1	-3.2	17.7	-10.7	-8.6	10.2	7.0	4.8
21. IMAGIN	3.4	12.6	10.4	9.8	60.2	3.5	4.6	32.9	25.0	20.3
22. PERFECT	4.8	13.8	11.7	11.1	64.8	5.0	5.9	35.2	26.7	21.8
23. TO WALES		9.3	7.0	6.6	44.1	0.2	1.5	25.8	19.7	15.9
24. THAT_THA	-9.3		-2.4	-2.5	14.4	-9.0	-7.3	9.0	6.4	4.5
25. SECONDAR	-7.0	2.4		-0.1	21.4	-6.7	-5.1	13.3	9.9	7.5
26. WOULD ON	-6.6	2.5	0.1		20.4	-6.3	-4.7	12.9	9.7	7.4
27. TELL	-44.1	-14.4	-21.4	-20.4		-42.6	-33.4	-3.7	-5.9	-7.7
28. COLOUR	-0.2	9.0	6.7	6.3	42.6		1.3	25.1	19.3	15.6
29. STOOD	-1.5	7.3	5.1	4.7	33.4	-1.3		20.9	16.2	13.1
30. INT1	-25.8	-9.0	-13.3	-12.9	3.7	-25.1	-20.9		-2.1	-3.8
31. INT2	-19.7	-6.4	-9.9	-9.7	5.9	-19.3	-16.2	2.1		-1.7
32. INT3	-15.9	-4.5	-7.5	-7.4	7.7	-15.6	-13.1	3.8	1.7	

Not only are there fewer combinations of estimates that reveal significant differences between them, but when the tokens are ranked in descending order of severity, it emerges that virtually no adjacent tokens differ significantly from one another, making it difficult to divide them into more than three clusters. This is plotted in Table 3.12.

Table 3.12. Ranking of tokens by clusters of adjacent severity scores (RP version only).

Token	Estimate	Error Category
(1) PERFECT	3.778	Stress
(1) IMAGIN	3.592	Stress
(1) THIN	3.468	Phonemic
(1) AUTHOR	3.250	Phonemic
(1) TO_WALES	3.152	Suprasegmental
(1) COLOUR	3.125	Phonemic
(1) RED	3.093	Realisational
(1) BED	3.018	Phonemic
(1) BAT	3.004	Phonemic
(1) FILM	2.946	Distributional
(1) VAN	2.938	Phonemic
(1) STOOD	2.918	Phonemic
(1) WINE	2.713	Phonemic
(1) DEAD	2.603	Realisational
(1) TIE	2.300	Realisational
(1) BREATHE	2.265	Phonemic
(1) WOULD_ON	2.030	Suprasegmental
(1) SECONDAR	2.002	Suprasegmental
(1) CAR	1.736	Distributional
(1) WEATHER	1.642	Phonemic
(1) BOTH	1.582	Phonemic
(1) THAT_THA	1.517	Suprasegmental
(1) NEW	1.457	Distributional
(1) HOT_TEA	1.400	Distributional
(1) ICE	1.252	Realisational
(1) OFF	0.981	Phonemic
(1) THAT	0.929	Phonemic
(1) INT3	0.657	Suprasegmental
(1) INDIA	0.620	Distractor
(1) INT2	0.402	Suprasegmental

(2) INT1	0.166	Suprasegmental
(3) TELL	3.129e-17	Realisational

3.2.5 Overall assessment of individual tokens in the GA version

The same analyses were applied to the severity scores in the GA version of the experiment. The severity estimates are presented in Table 3.13, and are plotted in Figure 3.9 below.

Table 3.13. Overall severity estimates for all 32 tokens (GA version only)

Token	Estimate	Standard Error
BED	3.660	0.100
BAT	3.088	0.125
VAN	3.546	0.109
WINE	2.828	0.121
THIN	3.233	0.102
AUTHOR	3.305	0.098
BOTH	3.151	0.115
OFF	1.382	0.120
THAT	1.817	0.132
WEATHER	3.225	0.102
BREATHE	2.513	0.159
RED	3.665	0.108
ICE	0.836	0.103
TIE	0.988	0.122
DEAD	3.465	0.114
FILM	3.172	0.115
CAR	2.428	0.117
HOT_TEA	1.417	0.139
INDIA	0.776	0.036
NEW	0.205	0.040
IMAGIN	3.827	0.078
PERFECT	3.871	0.069
TO_WALES	3.668	0.074
THAT_THA	1.063	0.120
SECONDARY	1.821	0.119
WOULD_ON	0.723	0.103
TELL	0.690	0.087
COLOUR	0.995	0.111
STOOD	1.004	0.113
INT1	0.082	0.043
INT2	0.639	0.091
INT3	0.801	0.104

On the whole, the highest and lowest estimates follow the same pattern as in the RP version of the experiment, but the different position of certain individual tokens (notably BOTH and TIE) is striking. It is also interesting to note that NEW has replaced TELL as the lowest estimate for a token that does not involve intonation. However, as is clear from Tables 3.14 to 3.16, there are even fewer significant differences between tokens in the GA version of the experiment than in the RP version. This will render attempts to rank these tokens in descending order of severity less reliable.

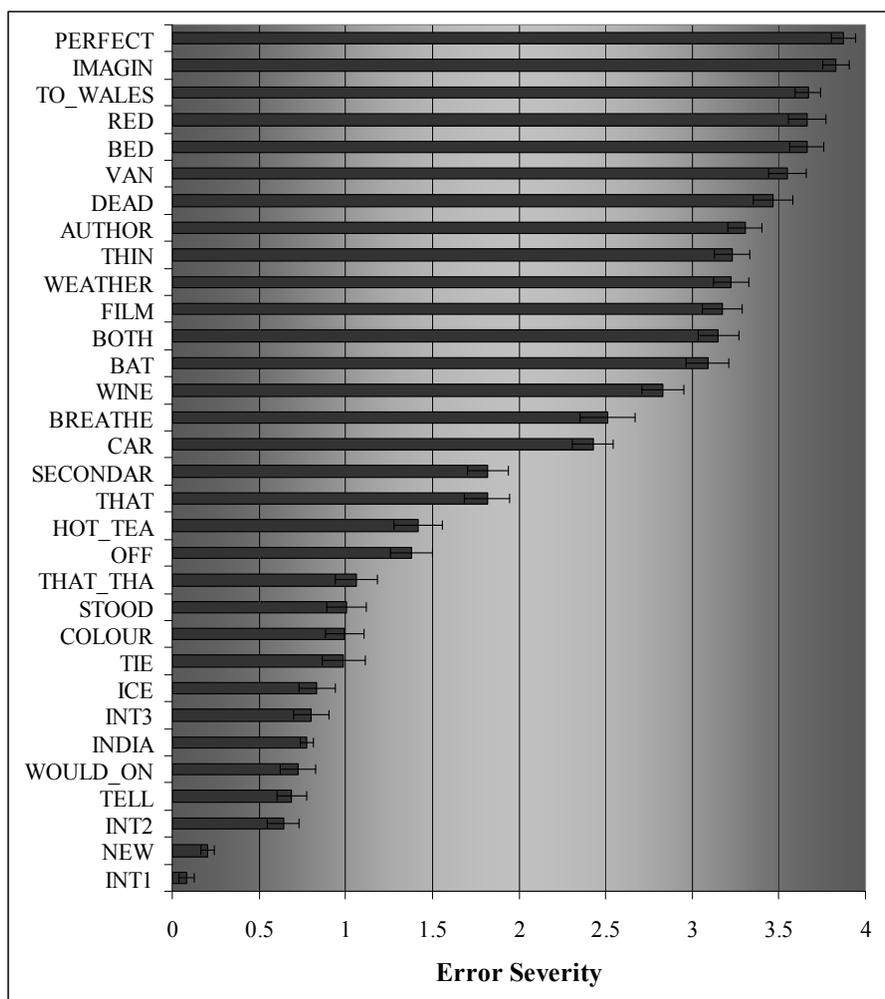


Figure 3.9. Bar chart with error bars showing overall severity estimates and standard errors for all 32 tokens (GA version).

For the GA version, the Wald Z score was calculated to establish if there were any significant differences between estimates. In this form of the experiment, significance ($Z \geq |2|$) was obtained in even fewer cases than in the RP version (76% as opposed to 82%). Where *no* significant difference between any two estimates exists, this is highlighted in **bold** (see Tables 3.14 to 3.16).

Table 3.14. Wald Z scores for all possible combinations of severity estimates in the GA version of the experiment (tokens 1–11).

	1. BED	2. BAT	3. VAN	4. WINE	5. THIN	6. AUTHOR	7. BOTH	8. OFF	9. THAT	10. WEATHER	11. BREATHE
1. BED		2.5	0.5	3.8	2.1	1.8	2.4	10.3	7.9	2.2	4.4
2. BAT	-2.5		-2.0	1.1	-0.6	-1.0	-0.3	7.0	5.0	-0.6	2.0
3. VAN	-0.5	2.0		3.1	1.5	1.2	1.8	9.5	7.2	1.5	3.9
4. WINE	-3.8	-1.1	-3.1		-1.8	-2.2	-1.4	6.0	4.0	-1.8	1.1
5. THIN	-2.1	0.6	-1.5	1.8		-0.4	0.4	8.3	6.1	0.0	2.8
6. AUTHOR	-1.8	1.0	-1.2	2.2	0.4		0.7	8.8	6.5	0.4	3.1
7. BOTH	-2.4	0.3	-1.8	1.4	-0.4	-0.7		7.5	5.4	-0.3	2.3
8. OFF	-10.3	-7.0	-9.5	-6.0	-8.3	-8.8	-7.5		-1.7	-8.3	-4.1
9. THAT	-7.9	-5.0	-7.2	-4.0	-6.1	-6.5	-5.4	1.7		-6.0	-2.4
10. WEATHER	-2.2	0.6	-1.5	1.8	0.0	-0.4	0.3	8.3	6.0		2.7
11. BREATHE	-4.4	-2.0	-3.9	-1.1	-2.8	-3.1	-2.3	4.1	2.4	-2.7	
12. RED	0.0	2.5	0.6	3.7	2.1	1.8	2.3	10.0	7.7	2.1	4.3
13. ICE	-13.9	-9.9	-12.8	-8.9	-11.7	-12.3	-10.6	-2.4	-4.2	-11.6	-6.4
14. TIE	-12.0	-8.5	-11.1	-7.6	-10.0	-10.6	-9.1	-1.6	-3.3	-10.0	-5.4
15. DEAD	-0.9	1.6	-0.4	2.7	1.1	0.8	1.4	8.9	6.7	1.1	3.5
16. FILM	-2.3	0.4	-1.7	1.5	-0.3	-0.6	0.1	7.6	5.5	-0.2	2.4
17. CAR	-5.7	-2.7	-4.9	-1.7	-3.7	-4.1	-3.1	4.4	2.5	-3.6	-0.3
18. HOT TEA	-9.4	-6.3	-8.6	-5.4	-7.5	-8.0	-6.8	0.1	-1.5	-7.5	-3.7
19. INDIA	-21.2	-14.4	-19.2	-13.1	-17.8	-19.0	-15.8	-3.9	-6.2	-17.8	-8.9
20. NEW	-24.7	-17.5	-22.6	-16.4	-21.4	-22.6	-19.1	-7.4	-9.4	-21.4	-11.6
21. IMAGIN	0.9	3.6	1.5	5.0	3.3	3.0	3.5	12.3	9.6	3.3	5.5
22. PERFECT	1.2	4.0	1.8	5.5	3.7	3.4	3.9	13.2	10.3	3.8	6.0
23. TO WALES	0.0	2.9	0.7	4.3	2.5	2.1	2.7	11.8	9.0	2.5	5.0
24. THAT THA	-11.7	-8.2	-10.8	-7.3	-9.7	-10.3	-8.9	-1.3	-3.0	-9.7	-5.2
25. SECONDAR	-8.4	-5.2	-7.6	-4.2	-6.4	-6.8	-5.7	1.8	0.0	-6.4	-2.5
26. WOULD ON	-14.4	-10.4	-13.3	-9.4	-12.2	-12.9	-11.2	-3.0	-4.7	-12.2	-6.8
27. TELL	-15.8	-11.3	-14.6	-10.3	-13.4	-14.2	-12.2	-3.3	-5.2	-13.4	-7.4
28. COLOUR	-12.6	-8.9	-11.6	-7.9	-10.5	-11.1	-9.6	-1.7	-3.4	-10.5	-5.6
29. STOOD	-12.4	-8.8	-11.5	-7.8	-10.4	-10.9	-9.4	-1.6	-3.3	-10.3	-5.5
30. INT1	-25.0	-17.9	-22.9	-16.8	-21.7	-23.0	-19.5	-8.0	-10.0	-21.7	-12.0
31. INT2	-15.8	-11.3	-14.6	-10.3	-13.4	-14.1	-12.2	-3.5	-5.3	-13.4	-7.5
32. INT3	-14.0	-10.0	-12.9	-9.0	-11.8	-12.4	-10.7	-2.6	-4.3	-11.8	-6.5

Table 3.15. Wald Z scores for all possible combinations of severity estimates in the GA version of the experiment (tokens 12–22).

	12. RED	13. ICE	14. TIE	15. DEAD	16. FILM	17. CAR	18. HOT_TEA	19. INDIA	20. NEW	21. IMAGIN	22. PERFECT
1. BED	0.0	13.9	12.0	0.9	2.3	5.7	9.4	21.2	24.7	-0.9	-1.2
2. BAT	-2.5	9.9	8.5	-1.6	-0.4	2.7	6.3	14.4	17.5	-3.6	-4.0
3. VAN	-0.6	12.8	11.1	0.4	1.7	4.9	8.6	19.2	22.6	-1.5	-1.8
4. WINE	-3.7	8.9	7.6	-2.7	-1.5	1.7	5.4	13.1	16.4	-5.0	-5.5
5. THIN	-2.1	11.7	10.0	-1.1	0.3	3.7	7.5	17.8	21.4	-3.3	-3.7
6. AUTHOR	-1.8	12.3	10.6	-0.8	0.6	4.1	8.0	19.0	22.6	-3.0	-3.4
7. BOTH	-2.3	10.6	9.1	-1.4	-0.1	3.1	6.8	15.8	19.1	-3.5	-3.9
8. OFF	-10.0	2.4	1.6	-8.9	-7.6	-4.4	-0.1	3.9	7.4	-12.3	-13.2
9. THAT	-7.7	4.2	3.3	-6.7	-5.5	-2.5	1.5	6.2	9.4	-9.6	-10.3
10. WEATHER	-2.1	11.6	10.0	-1.1	0.2	3.6	7.5	17.8	21.4	-3.3	-3.8
11. BREATHE	-4.3	6.4	5.4	-3.5	-2.4	0.3	3.7	8.9	11.6	-5.5	-6.0
12. RED		13.4	11.7	0.9	2.2	5.5	9.1	20.2	23.5	-0.9	-1.2
13. ICE	-13.4		-0.7	-12.1	-10.7	-7.2	-2.4	0.4	4.4	-16.5	-17.6
14. TIE	-11.7	0.7		-10.5	-9.2	-6.0	-1.6	1.3	4.9	-14.2	-15.1
15. DEAD	-0.9	12.1	10.5		1.3	4.5	8.1	17.9	21.2	-1.9	-2.2
16. FILM	-2.2	10.7	9.2	-1.3		3.2	6.9	15.9	19.2	-3.4	-3.8
17. CAR	-5.5	7.2	6.0	-4.5	-3.2		3.9	10.8	14.2	-7.2	-7.8
18. HOT TEA	-9.1	2.4	1.6	-8.1	-6.9	-3.9		3.7	6.8	-11.1	-11.8
19. INDIA	-20.2	-0.4	-1.3	-17.9	-15.9	-10.8	-3.7		7.6	-26.8	-29.6
20. NEW	-23.5	-4.4	-4.9	-21.2	-19.2	-14.2	-6.8	-7.6		-30.8	-33.9
21. IMAGIN	0.9	16.5	14.2	1.9	3.4	7.2	11.1	26.8	30.8		-0.3
22. PERFECT	1.2	17.6	15.1	2.2	3.8	7.8	11.8	29.6	33.9	0.3	
23. TO WALES	0.0	15.9	13.7	1.1	2.6	6.5	10.6	26.3	30.5	-1.0	-1.4
24. THAT THA	-11.4	1.0	0.3	-10.2	-8.9	-5.7	-1.4	1.8	5.3	-13.9	-14.8
25. SECONDAR	-8.1	4.4	3.5	-7.0	-5.8	-2.6	1.6	6.7	10.2	-10.2	-10.9
26. WOULD ON	-14.0	-0.5	-1.2	-12.6	-11.2	-7.7	-2.9	-0.4	3.6	-17.1	-18.3
27. TELL	-15.3	-0.8	-1.4	-13.8	-12.3	-8.5	-3.2	-0.7	3.8	-19.0	-20.4
28. COLOUR	-12.2	0.7	0.0	-11.0	-9.7	-6.3	-1.7	1.5	5.3	-15.0	-16.1
29. STOOD	-12.1	0.8	0.1	-10.8	-9.5	-6.2	-1.6	1.5	5.2	-14.8	-15.8
30. INT1	-23.8	-5.2	-5.5	-21.5	-19.6	-14.6	-7.3	-8.8	-1.5	-31.0	-34.0
31. INT2	-15.2	-1.0	-1.6	-13.8	-12.3	-8.6	-3.4	-1.1	3.3	-18.8	-20.2
32. INT3	-13.5	-0.2	-0.8	-12.2	-10.8	-7.3	-2.5	0.2	4.1	-16.6	-17.8

Table 3.16. Wald Z scores for all possible combinations of severity estimates in the GA version of the experiment (tokens 23–32).

	23. TO_WALES	24. THAT_THA	25. SECONDAR	26. WOULD_ON	27. TELL	28. COLOUR	29. STOOD	30. INT1	31. INT2	32. INT3
1. BED	0.0	11.7	8.4	14.4	15.8	12.6	12.4	25.0	15.8	14.0
2. BAT	-2.9	8.2	5.2	10.4	11.3	8.9	8.8	17.9	11.3	10.0
3. VAN	-0.7	10.8	7.6	13.3	14.6	11.6	11.5	22.9	14.6	12.9
4. WINE	-4.3	7.3	4.2	9.4	10.3	7.9	7.8	16.8	10.3	9.0
5. THIN	-2.5	9.7	6.4	12.2	13.4	10.5	10.4	21.7	13.4	11.8
6. AUTHOR	-2.1	10.3	6.8	12.9	14.2	11.1	10.9	23.0	14.1	12.4
7. BOTH	-2.7	8.9	5.7	11.2	12.2	9.6	9.4	19.5	12.2	10.7
8. OFF	-11.8	1.3	-1.8	3.0	3.3	1.7	1.6	8.0	3.5	2.6
9. THAT	-9.0	3.0	0.0	4.7	5.2	3.4	3.3	10.0	5.3	4.3
10. WEATHER	-2.5	9.7	6.4	12.2	13.4	10.5	10.3	21.7	13.4	11.8
11. BREATHE	-5.0	5.2	2.5	6.8	7.4	5.6	5.5	12.0	7.5	6.5
12. RED	0.0	11.4	8.1	14.0	15.3	12.2	12.1	23.8	15.2	13.5
13. ICE	-15.9	-1.0	-4.4	0.5	0.8	-0.7	-0.8	5.2	1.0	0.2
14. TIE	-13.7	-0.3	-3.5	1.2	1.4	0.0	-0.1	5.5	1.6	0.8
15. DEAD	-1.1	10.2	7.0	12.6	13.8	11.0	10.8	21.5	13.8	12.2
16. FILM	-2.6	8.9	5.8	11.2	12.3	9.7	9.5	19.6	12.3	10.8
17. CAR	-6.5	5.7	2.6	7.7	8.5	6.3	6.2	14.6	8.6	7.3
18. HOT TEA	-10.6	1.4	-1.6	2.9	3.2	1.7	1.6	7.3	3.4	2.5
19. INDIA	-26.3	-1.8	-6.7	0.4	0.7	-1.5	-1.5	8.8	1.1	-0.2
20. NEW	-30.5	-5.3	-10.2	-3.6	-3.8	-5.3	-5.2	1.5	-3.3	-4.1
21. IMAGIN	1.0	13.9	10.2	17.1	19.0	15.0	14.8	31.0	18.8	16.6
22. PERFECT	1.4	14.8	10.9	18.3	20.4	16.1	15.8	34.0	20.2	17.8
23. TO WALES		13.4	9.6	16.6	18.5	14.5	14.2	30.7	18.3	16.1
24. THAT_THA	-13.4		-3.2	1.5	1.8	0.3	0.3	6.0	2.0	1.2
25. SECONDAR	-9.6	3.2		4.9	5.5	3.6	3.5	10.7	5.6	4.6
26. WOULD ON	-16.6	-1.5	-4.9		0.2	-1.3	-1.3	4.4	0.4	-0.4
27. TELL	-18.5	-1.8	-5.5	-0.2		-1.5	-1.6	4.7	0.3	-0.6
28. COLOUR	-14.5	-0.3	-3.6	1.3	1.5		0.0	6.0	1.8	0.9
29. STOOD	-14.2	-0.3	-3.5	1.3	1.6	0.0		5.9	1.8	0.9
30. INT1	-30.7	-6.0	-10.7	-4.4	-4.7	-6.0	-5.9		-4.2	-4.9
31. INT2	-18.3	-2.0	-5.6	-0.4	-0.3	-1.8	-1.8	4.2		-0.8
32. INT3	-16.1	-1.2	-4.6	0.4	0.6	-0.9	-0.9	4.9	0.8	

In the GA version, there are even fewer combinations of estimates that are significantly different – which also goes for most adjacent tokens. As in the RP version, this results in no more than three clusters (see Table 3.17), but here these are clearly divided into: an “upper” range (> 2.2); an “intermediate” range (2.2 to 0.4); and a “lower” range (< 0.4). Once again, the eccentric position of NEW is particularly salient.

Table 3.17. Ranking of tokens by clusters of adjacent severity scores (GA version only).

Token	Estimate	Error Category
(1) PERFECT	3.871	Stress
(1) IMAGIN	3.827	Stress
(1) TO_WALES	3.668	Suprasegmental
(1) RED	3.665	Realisational
(1) BED	3.660	Phonemic
(1) VAN	3.546	Phonemic
(1) DEAD	3.465	Realisational
(1) AUTHOR	3.305	Phonemic
(1) THIN	3.233	Phonemic
(1) WEATHER	3.225	Phonemic
(1) FILM	3.172	Distributional
(1) BOTH	3.151	Phonemic
(1) BAT	3.088	Phonemic
(1) WINE	2.828	Phonemic
(1) BREATHE	2.513	Phonemic
(1) CAR	2.428	Distributional
(2) SECONDAR	1.821	Suprasegmental
(2) THAT	1.817	Phonemic
(2) HOT_TEA	1.417	Distributional
(2) OFF	1.382	Phonemic
(2) THAT_THA	1.063	Suprasegmental
(2) STOOD	1.004	Phonemic
(2) COLOUR	0.995	Phonemic
(2) TIE	0.988	Realisational
(2) ICE	0.836	Realisational
(2) INT3	0.801	Suprasegmental
(2) INDIA	0.776	Distractor
(2) WOULD_ON	0.723	Suprasegmental

(2) TELL	0.690	Realisational
(2) INT2	0.639	Suprasegmental
(3) NEW	0.205	Distributional
(3) INT1	0.082	Suprasegmental

3.2.6 Overall assessment of individual tokens: differences between versions

It would seem that there are too few clusters of severity estimates in either the RP or the GA version to establish a detailed hierarchy of error for each of these versions (but see 5.2.2). It is, however, possible to discuss differences between the two versions by comparing and contrasting the relevant estimates for each token. The MLwiN program was used to calculate these differences between versions. A negative result means that this token was judged more severely in the GA version, whereas a positive result suggests a more severe assessment in the RP version. These results are significant at $p < .05$ ($\chi^2 > 3.8$, $df = 1$), unless listed as “n.s.” in Table 3.18.

Table 3.18. Overall severity by token: differences between RP and GA versions.

Token	Difference RP ~ GA	χ^2 (df = 1)	Sig.
BED	-0.64	34.02	
BAT	-0.08	0.42	n.s.
VAN	-0.61	25.82	
WINE	-0.12	0.82	n.s.
THIN	0.23	5.00	
AUTHOR	-0.06	0.30	n.s.
BOTH	-1.57	146.65	
OFF	-0.4	10.16	
THAT	-0.89	42.11	
WEATHER	-1.58	179.01	
BREATHE	-0.25	2.21	n.s.
RED	-0.57	25.74	
ICE	0.42	13.48	
TIE	1.31	103.25	

DEAD	-0.86	44.79	
FILM	-0.23	3.38	n.s.
CAR	-0.69	30.33	
HOT_TEA	-0.02	0.01	n.s.
INDIA	-0.16	15.46	
NEW	1.25	223.78	
IMAGIN	-0.24	8.31	
PERFECT	-0.09	1.58	n.s.
TO_WALES	-0.52	37.28	
THAT_THA	0.45	11.94	
SECONDAR	0.18	2.070	n.s.
WOULD_ON	1.31	123.48	
TELL	-0.69	62.60	
COLOUR	2.13	356.07	
STOOD	1.91	260.57	
INT1	0.08	2.71	n.s.
INT2	-0.24	6.26	
INT3	-0.14	1.69	n.s.

Table 3.18 shows that there were significant differences between the two versions in no fewer than 22 cases (out of 32). It also reveals a high level of inter-version similarity for four potential phonemic errors (BAT, WINE, AUTHOR and BREATHE), two distributional/realisational differences (FILM and HOT_TEA), one stress error (PERFECT) and three suprasegmental phenomena (SECONDAR,

INT1 and INT3). Of these, it is only HOT_TEA and the three suprasegmental phenomena that fall outside the “upper” and “upper intermediate” ranges (> 2.0) of the overall hierarchy of error. Interestingly, this suggests that all judges taking part in the experiment considered a potential error such as BAT to be equally serious, no matter which version of the experiment they had opted to take. The same consistently high severity across versions is associated with WINE, AUTHOR, BREATHE, FILM and PERFECT.

It is also interesting to note that the estimates for SECONDAR were consistent between the two versions. This “mirrored token” (see 2.1.3 and 2.3) featured the four-syllable GA pronunciation of *secondary* in the RP version as the potential error, whereas in the GA version it was the weakened RP pronunciation of this word that had served this purpose. The result shows that this kind of cross-Atlantic token mirroring does not necessarily have to affect the severity of the judges. It also suggests that the RP pronunciation of *secondary* is as unproblematic to judges of GA as the GA pronunciation is to judges of RP. This is, however, demonstrably untrue of other examples of token mirroring such as CAR and NEW. In these cases, however, the potential errors are associated not only with the other main cross-Atlantic variety of English, but also with what would appear to be stigmatised local pronunciations.

There was much less inter-version consistency for the other 22 tokens, notably in those cases where the difference between the two severity estimates for each token exceeded one Likert scale point. In the RP version, such dramatically increased severity is associated with: two potential phonemic errors (COLOUR and STOOD); two realisational/distributional differences (TIE and NEW); and one suprasegmental feature (WOULD_ON). In the GA version, on the other hand, such increases were attested for two phoneme contrasts BOTH and WEATHER – with phonemic THAT (-0.89) and realisational/distributional DEAD (-0.86) just below 1.0 (see also Figure 3.10).

As will be shown in the token-by-token discussion of the severity scores, some of these differences may be explained in terms of accent variation and stigmatised local features. Variation and stigmatisation may also account for the apparent lack of inter-version consistency as regards the six tokens featuring either /θ/ or /ð/. While AUTHOR (medial /θ/) and BREATHE (final /ð/) were rated similarly, WEATHER (medial /ð/), BOTH (final /θ/), and THAT (initial /ð/) were considered to be more serious errors in the GA version, whereas THIN (initial /θ/) was judged more severely in the RP version. However, the position of the relevant phoneme in the word and the relative frequency of the word in question in the language may also account for this outcome. Finally, there could perhaps have been subtle differences in performance between the two actors that overemphasised certain features in one of the two versions.

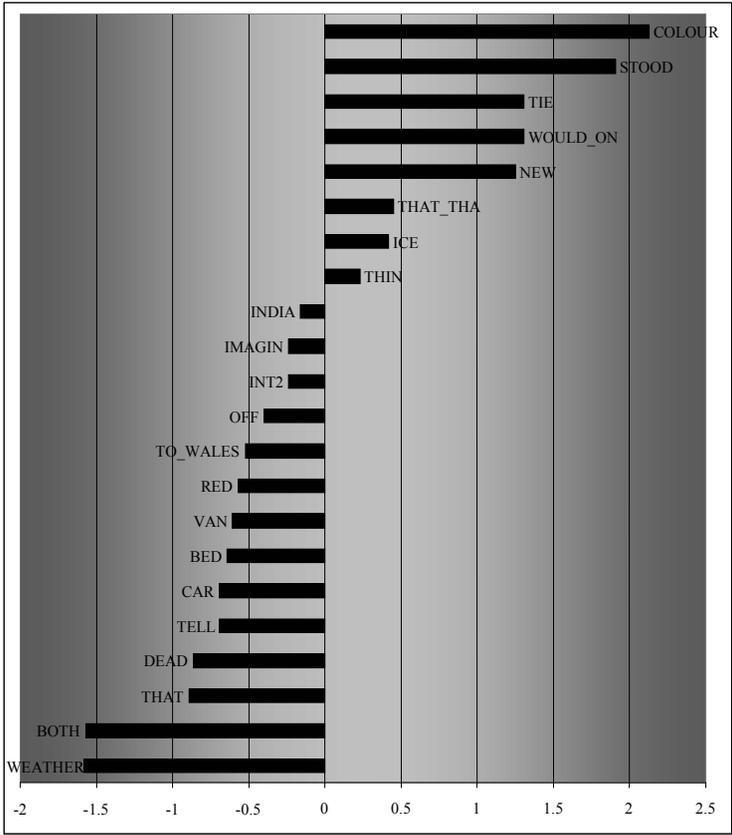


Figure 3.10. Significant differences in severity (by token) between the GA and RP versions of the experiment. Negative values represent increased severity in the GA version, positive values increased severity in the RP version.

3.3 Assessment of individual tokens by age and sex

3.3.1 General overview of the results

By means of multi-level modelling, it is possible to calculate the effects of sex and age on the severity assessment of individual tokens (see 3.3.2 and 3.3.3). This procedure reveals that female respondents tend to judge BED, HOT_TEA and NEW significantly more leniently, and that older respondents (of either sex) are significantly more tolerant of WINE, THIN, AUTHOR, IMAGIN, THAT_THA, WOULD_ON and INT2. In addition, both the female judges and the older judges (of either sex) tend to give significantly lower severity scores to ICE, TIE, CAR,

COLOUR and STOOD. Some of these tokens also have significantly lower estimates in the GA version of the experiment.

3.3.2 Assessment of individual tokens by sex

For each token, the difference in severity scores between male and female respondents has been represented in Table 3.19. Significant differences (if $Z \geq |2|$) have been highlighted in **bold**. A negative value means that female respondents judged the token less severely; a positive value indicates greater leniency on the part of the male respondents. Since the only significant values are negative, this suggests that no token was judged significantly more leniently by males. These findings correspond with the consistently lower overall severity scores for women.

Table 3.19. Differences in severity between male and female respondents (centralised for age).

Token	Sex Diff.	Standard Error	Wald Z
BED	0.126	0.134	-3.9
BAT	-0.130	0.160	-1.2
VAN	-0.100	0.109	-1.2
WINE	-0.490	0.114	-0.2
THIN	-0.350	0.137	-0.3
AUTHOR	-0.230	0.134	-1.6
BOTH	-0.070	0.121	1.6
OFF	-0.390	0.126	-0.0
THAT	-0.480	0.140	0.6
WEATHER	0.024	0.040	0.9
BREATHE	-0.250	0.054	-0.8
RED	-0.040	0.079	-0.9
ICE	-0.110	0.073	-4.4
TIE	-0.060	0.086	-2.6

DEAD	-0.100	0.131	-1.7
FILM	0.024	0.121	-0.5
CAR	-0.230	0.129	-3.0
HOT_TEA	0.029	0.030	-3.4
DISTRACT	-0.350	0.114	0.6
NEW	-0.470	0.133	-4.6
IMAGIN	0.007	0.050	-0.6
PERFECT	0.007	0.090	-1.5
TO_WALES	0.042	0.108	-0.8
THAT_THA	0.126	0.134	-0.8
SECONDAR	-0.130	0.160	0.2
WOULD_ON	-0.100	0.109	-1.8
TELL	-0.490	0.114	0.9
COLOUR	-0.350	0.137	-3.1
STOOD	-0.230	0.134	-3.5
INT1	-0.070	0.121	0.1
INT2	-0.390	0.126	0.1
INT3	-0.480	0.140	0.4

Once the estimates (and the corresponding standard errors) for each token have been calculated for the male respondents, the estimates for female respondents can be found by adding the value “SEX DIFF” to the estimates for men. This has been done (in those cases where the differences are significant) for Table 3.20, which shows that the tokens concerned are all segmental. This implies that, among other things, there is general agreement among male and female participants concerning stress and suprasegmental features. These differences are also plotted in Figure 3.11.

Table 3.20. Severity estimates for male and female respondents for each token where the difference is significant (centralised for age).

Token	Const. (M)	Standard Error	Estim. (F)	Error Category
BED	3.275	0.079	2.841	Phonemic
ICE	1.086	0.080	0.592	Realisational
TIE	1.929	0.094	1.578	Realisational
CAR	1.981	0.088	1.592	Distributional
HOT_TEA	1.406	0.098	0.928	Distributional
NEW	0.312	0.040	0.064	Distributional
COLOUR	2.874	0.077	2.519	Phonemic
STOOD	2.514	0.091	2.044	Phonemic

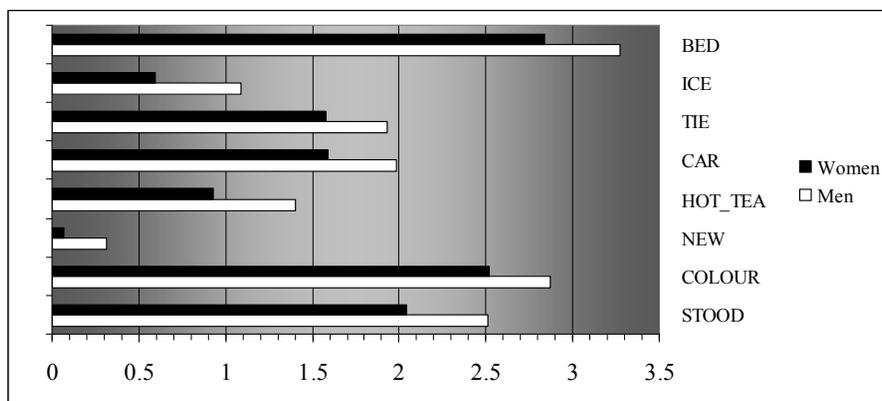


Figure 3.11. Severity estimates for male and female respondents for each token where the difference is significant (centralised for age).

It could be argued that some of the differences between estimates for male and female respondents are in fact the effects of variables other than sex. After all, 55.1% of judges in the RP version of the experiment are men, as opposed to 45% in the GA form. If female respondents are on the whole more tolerant of tokens that were also judged less severely in the GA version, this could possibly be attributed to their relative overrepresentation in this group. But even though this may be true for ICE, TIE, NEW, COLOUR and STOOD, it should be also noted that BED and CAR were in fact judged more leniently in the RP form. This would suggest that the differences for male and female participants cannot be completely accounted for in terms of unequal distributions of men and women taking part in the two versions. The implications of the dissimilarities between these two groups of respondents will be further discussed in 3.4.4, 3.5 and 3.7.

3.3.3 Assessment of individual tokens by age

The regression coefficient for age (after centralisation) turned out to be significant for a number of tokens, as can be observed in Table 3.21 below. A negative coefficient indicates that an increase in age results in increased error tolerance, whereas a positive coefficient represents an increase in severity. Since the only significant values are negative, this suggests that no token was judged significantly more leniently by younger respondents. These findings correspond with the consistently lower overall severity assessment score for older respondents.

This means that, as in the case of sex, the effect of age on token assessment cannot simply be explained on the basis of the proportional overrepresentation of older judges in the GA version (39.2% of GA judges were over 40, as opposed to only 25.7% of RP judges). Even if older respondents and women are both more likely to be tolerant of potential errors such as ICE, TIE, CAR, COLOUR and STOOD, this must be at least partially ascribed to the independent effects of age and sex. The effects of age on token assessment are also plotted in Figure 3.12 below (for those tokens which have significant age coefficients). These results will be analysed in more detail in 3.4.4, 3.5 and 3.7.

Table 3.21. Regression coefficients for token assessment by age (after centralisation). Significant coefficients have been highlighted in **bold** (if $Z \geq |2|$).

Token	Age	Standard Error	Wald Z
BED	-0.006	0.004	-1.5
BAT	-0.006	0.005	-1.3
VAN	-1e-03	0.004	-0.2
WINE	-0.014	0.005	-3.2
THIN	-0.026	0.004	-7.1
AUTHOR	-0.014	0.004	-4.0
BOTH	0.002	0.005	0.4
OFF	0.001	0.004	0.3
THAT	-0.003	0.005	-0.6
WEATHER	0.006	0.005	1.3
BREATHE	0.007	0.006	1.1
RED	-0.007	0.004	-1.7
ICE	-0.017	0.004	-4.2
TIE	-0.015	0.005	-3.0
DEAD	-0.005	0.005	-1.0
FILM	-0.006	0.004	-1.4
CAR	-0.013	0.005	-2.9
HOT_TEA	0.003	0.005	0.6
DISTRACT	0.002	0.001	1.3

NEW	-0.003	0.002	-1.7
IMAGIN	-0.011	0.003	-3.8
PERFECT	-0.003	0.003	-1.0
TO_WALES	-0.002	0.003	-0.6
THAT_THA	-0.024	0.005	-4.9
SECONDAR	-0.006	0.004	-1.3
WOULD_ON	-0.020	0.005	-4.3
TELL	8.45e-06	0.001	0
COLOUR	-0.024	0.004	-5.6
STOOD	-0.031	0.005	-6.2
INT1	-0.001	0.002	-0.8
INT2	-0.013	0.003	-3.8
INT3	-0.005	0.004	-1.3

Table 3.22. The effects of an increase of 25 years on token assessment

Token	Age (x25)	Error Categories
WINE	-0.358	Phonemic
THIN	-0.642	Phonemic
AUTHOR	-0.353	Phonemic
ICE	-0.431	Realisational
TIE	-0.374	Realisational
CAR	-0.335	Distributional
IMAGIN	-0.281	Stress
THAT_THA	-0.592	Suprasegmental
WOULD_ON	-0.509	Suprasegmental
COLOUR	-0.610	Phonemic
STOOD	-0.771	Phonemic
INT2	-0.318	Suprasegmental

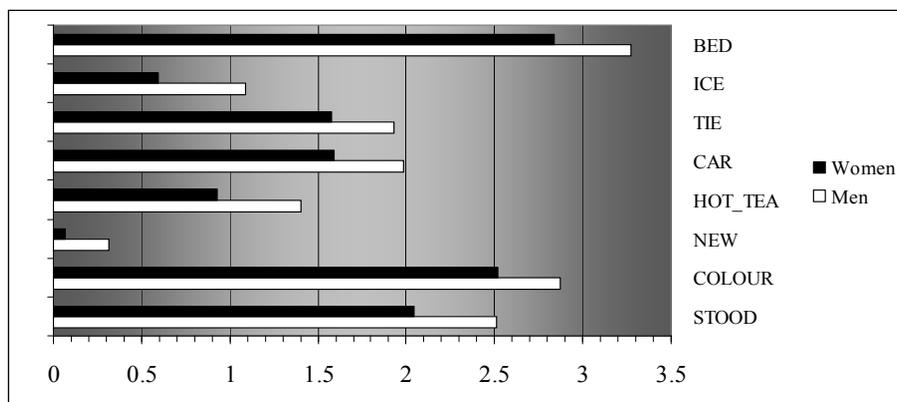


Figure 3.12. The effects of an increase of 25 years on token assessment in descending order of magnitude (for those tokens that have significant age coefficients).

3.4 Assessment of individual tokens by major accent group

3.4.1 Introduction and general overview of the results

In order to establish differences between the severity scores for individual tokens in each of the seven major accent groups, these were subjected to multi-

OFF					+							
THAT		+	+		+	+					+	?
WEATHER	+	+	+	+	+	+	?	?		+	+	+
BREATHE												
RED				+	+					+	?	
ICE						+						
TIE	+	+	+	+	+	+	+	+	+	+		+
DEAD	+	+	+			+				+	+	+
FILM							+	+	+			
CAR	+			+	+							
HOT_TEA												
DISTRACT		+										
NEW	+	+	+	+	+	+				+	+	+
IMAGIN												
PERFECT												
TO_WALES	+	+		+	+		+	+				
THAT_THA	+											
SECONDARY												
WOULD_ON	+	+	+	+	+	+	+	+	+	+	+	+
TELL	+	+	+	+	+		+	+		+	+	
COLOUR	+	+	+	+	+	+	+	+	+	+	+	+
STOOD	+	+	+	+	+	+	+	+	+	+	+	+
INT1												
INT2												
INT3												

While the results suggest that most of the differences may be explained by inter-version variation, the exceptions to this are either based on (a) inter-group variation within the RP form of the experiment, as is the case with *BED*, *VAN*, *OFF* and *FILM* or (b) potentially different levels of convergence for any of the three North American accent groups with any of the groups judging the RP version of the experiment. As is shown in the examples of *ICE* and *NEW*, such possible differences could conceivably point to additional inter-group variation within either the GA or the RP form of the experiment.

A case in point is *ICE*, which was judged similarly by all pairs of groups except for GB/NRP ~ CDN. Consequently, while *ICE* was not evaluated demonstrably differently by the three North American groups, the Canadians (unlike the two US groups) nevertheless had significantly different severity scores from one of the two British groups. Even if this could suggest a possible divergence of the CDN group from the US/GA and US/NGA groups, there is no direct statistical evidence to support this, since the various comparisons between North

American groups did not reveal any significant differences. As always, caution is required in interpreting null results since their exact causes are unknown – as well as being true null results, they could also be the effect of large sampling errors. But even if no evidence can be found to show that the CDN group judged this token differently from the US groups, one cannot exclude the possibility of inter-group variation in North America. Canadian Raising implies that a typically Canadian pronunciation of ICE is appreciably different from that of GA, and therefore it would not be surprising if CDN respondents were to judge this token differently.

Similarly, NEW was judged significantly differently by all cross-Atlantic and cross-Pacific pairs of accent groups, with the exception of the IRL judges. Of all the groups taking the RP form of the experiment, the latter are the only group not to show any significant differences from the North American judges for this specific token. Even though no differences were attested in any pairwise comparisons of the IRL judges with other groups within the RP version either (at least with regard to NEW), the fact that the IRL judges are the only group to agree with all other major accent groups on this would suggest that, at least as far as this token is concerned, the Irish respondents occupy an intermediate position between the other groups in the RP version and those in the GA form. Once again, however, there is no hard and fast evidence to prove this. In addition, it should be noted that, no matter how tempting it might be to generalise from this and claim an intermediate position for the IRL judges for all tokens, the examples of BOTH, TIE and STOOD show that such an assertion cannot be supported by the data. What the IRL evaluation of NEW does show, however, is that it cannot be assumed that the increased severity of groups within the RP form towards this token automatically extends to the Irish judges as well.

3.4.2 Pairwise comparisons between GB/RP, US/GA, GB/NRP and US/NGA

Since most differences between accent groups may be accounted for by inter-version variation, and much less frequently by inter-group variation, this will render it unnecessary to provide a detailed overview of the statistically significant differences for all 32 tokens in each of the 21 pairwise comparisons between accent groups. Instead, any salient differences will be discussed in detail in 3.5, which will provide a token-by-token analysis. However, an exception has been made for comparisons of GB/RP and US/GA, GB/RP and US/NGA, GB/NRP and US/GA, and GB/NRP and US/NGA, which will be treated below. Amongst other things, this is warranted by the fact that RP and GA are the most common pronunciation models for foreign learners. Firstly, it follows that it would be useful to know the differences in token evaluation between them, if only as a necessary corrective to Figure 3.10 (in 3.2.6). This table shows significant differences in severity (by token) between the GA and RP versions of the experiment, but not between the US/GA and GB/RP judges themselves. Secondly, it would be helpful to ascertain whether any of the tokens had been judged significantly differently by British speakers of RP vis-à-vis

those who speak a different variety, or by US speakers of GA as compared with those with a different accent. Since it is especially in Britain and the US that the judgements of native speakers of varieties other than RP and GA will be affected by their awareness of the prestige varieties, such comparisons will provide a much needed sociolinguistic perspective for error analysis (cf. 3.7). Even if a direct comparison between GB/RP and GB/NRP yields very few concrete results (see Table 3.23), and a comparison between US/GA and US/NGA none at all (see Table 3.24), indirect evidence for the differences between these varieties can be provided by instead contrasting GB/RP with US/NGA, GB/NRP with US/NGA and GB/NRP with US/NGA (see Figures 3.14 to 3.16).

Figure 3.13 shows that the significant differences between GB/RP and US/GA are a subset of those judged significantly differently in the RP and GA versions of the experiment (for the latter, see 3.2.6). While there appears to be a large amount of agreement between GB/RP and US/GA, it is striking that the GB/RP group attaches more importance to *THAT_THA* than the US/GA judges. No such result was found in any of the other three pairwise comparisons (see Figures 3.14 to 3.16), which all feature the same tokens on the “British” side (represented by positive values) but a varying number of different tokens on the “American” side (represented by negative values).

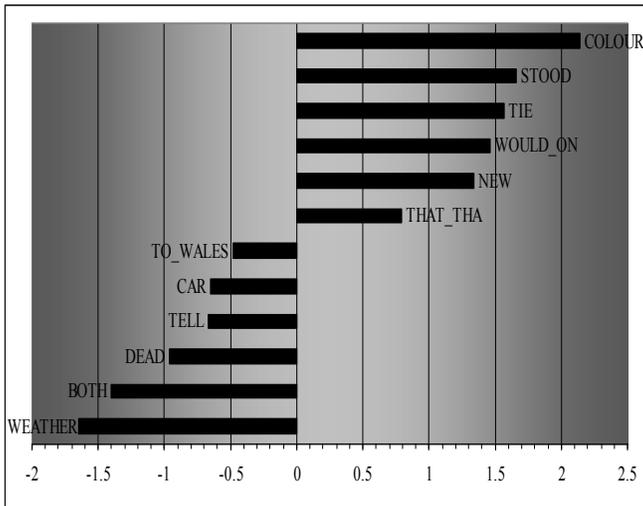


Figure 3.13. Significant differences in severity (by token) between the US/GA and GB/RP groups. Negative values represent increased severity in the US/GA group, positive values increased severity in the GB/RP group.

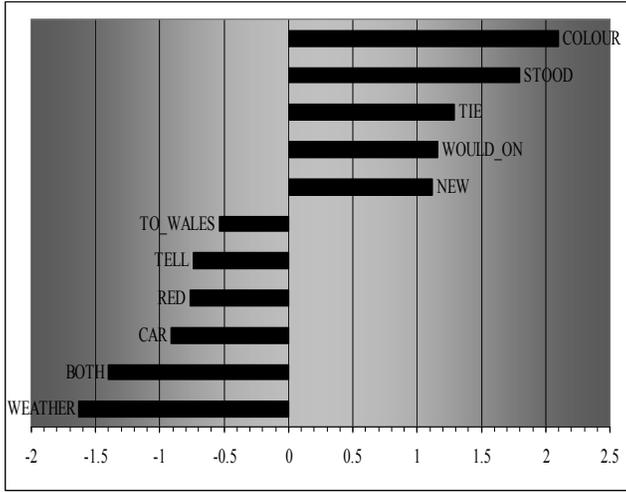


Figure 3.14 Significant differences in severity (by token) between the US/GA and GB/NRP groups. Negative values represent increased severity in the US/GA group; positive values increased severity in the GB/NRP group.

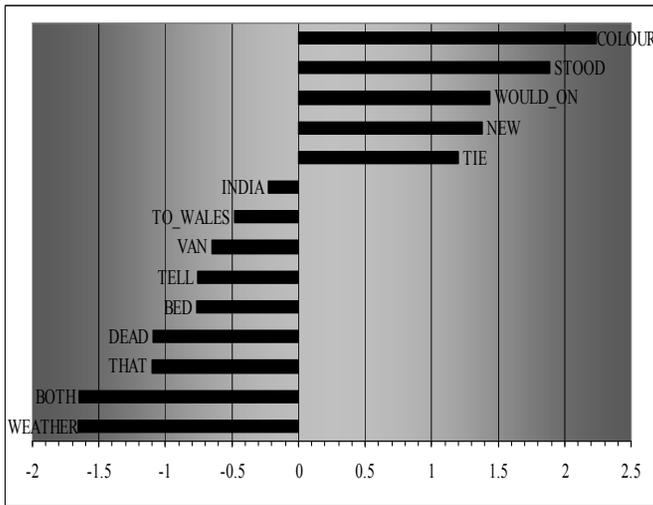


Figure 3.15. Significant differences in severity (by token) between the US/NGA and GB/RP groups. Negative values represent increased severity in the US/NGA group; positive values increased severity in the GB/RP group.

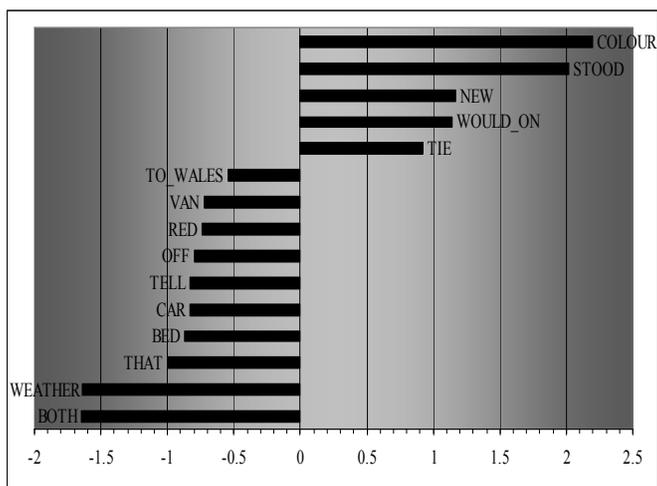


Figure 3.16. Significant differences in severity (by token) between the US/NGA and GB/NRP groups. Negative values represent increased severity in the US/NGA group; positive values increased severity in the GB/NRP group.

It is also interesting to note that the US/GA respondents, like the US/NGA group, tend to judge DEAD significantly less leniently than the GB/RP respondents (unlike the GB/NRP group). This may suggest that GB/RP judges, at least as compared to US/GA and/or US/NGA groups, tend to evaluate THAT_THA more strictly and DEAD less strictly than GB/NRP respondents – even though it should be noted that pairwise comparisons between GB/RP and GB/NRP are not statistically significant for these tokens (as can be seen from Table 3.23 in 3.4.1).

Similarly, while RED is judged more severely by the two American groups than by the GB/NRP speakers, this is not the case for the GB/RP speakers. Since uvular-*r* may indeed be found in some Northern British English but not in RP (see 3.5.8 and 4.2.7), it would make perfect sense if the GB/NRP group assessed RED more leniently than GB/RP speakers in a comparison with US/GA or US/NGA speakers. Nevertheless, it should be stressed that no statistically significant differences can be found between GB/NRP and GB/RP in this respect.

Finally, VAN, BED, OFF and THAT only appear in comparisons between US/NGA and the two British groups. This could conceivably be an indication that this group also attaches more importance to these tokens than the US/GA speakers. If it is true that these tokens are stigmatised or associated with a foreign accent, this would suggest that especially US/NGA speakers are less tolerant of such pronunciations. If these US/NGA speakers are to be perceived as deviant from the standard accent, their relative intolerance of these

pronunciations may be ascribed to greater “linguistic insecurity” in at least some speakers of minor accent groups (see 3.7).

3.4.3 Respondents’ comments on individual tokens

The extra comments volunteered by respondents can illustrate, or even help to explain, certain variations in the assessment of different tokens between major accent groups. Such observations sometimes reveal that particular pronunciations are stigmatised in certain communities, or are associated with different varieties of English. This may cause potential errors to be evaluated in dissimilar ways by respondents in different major accent groups. Space was provided in the online survey to provide for respondents’ comments on each token. This space was utilised in 17% of the 16,895 cases (see Figure 3.17).

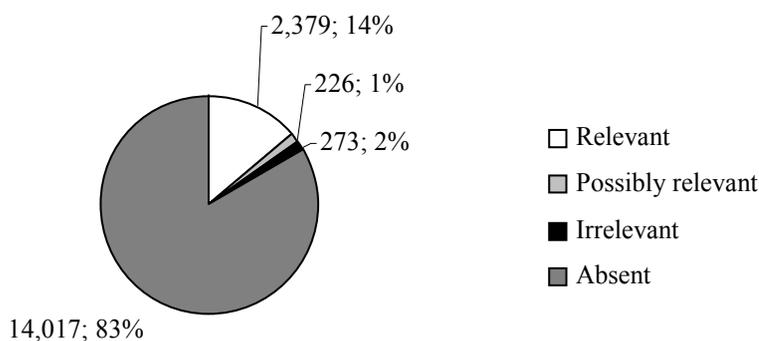


Figure 3.17. Respondents’ comments about individual tokens, divided into four categories according to relevance (n = 2,878).

The resulting comments can be subdivided as follows:

- (1) those *directly* relevant to the potential errors intended, or helped to identify these (14%);
- (2) those *possibly* relevant, for instance, if they referred to the words containing the intended errors (1%);
- (3) “irrelevant” comments which, although sometimes useful and revealing, did not refer to the intended errors, or help to identify these (2%).

Since only 273 comments were categorised as “irrelevant” (9.5% of the total number of actual comments), this implies that the overwhelming majority of

comments were relevant. These comments are clearly useful for a discussion of the assessment of individual tokens.

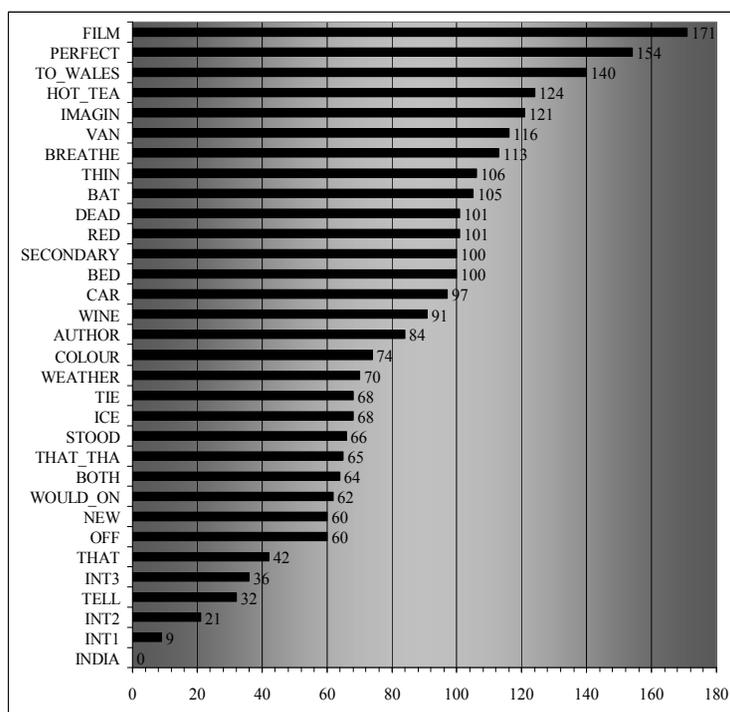


Figure 3.18. Numbers of relevant or possibly relevant comments broken down by token.

If the numbers of relevant and possibly relevant comments are broken down by token, as in Figure 3.18, it becomes apparent that certain tokens generated more observations than others: while FILM, for instance, inspired 171 relevant or possibly relevant comments, THAT only gave rise to 42. What consequences, if any, this may have is discussed in Section 3.5.

The 2,605 relevant or possibly relevant comments were further subdivided according to the following two criteria: (1) *tone* of the observations, and (2) *subject matter*. Whilst the tone of comments may illustrate the possible affect produced by tokens, and – where negative – attest to possible stigmatisation, their subject matter may amongst other things indicate with which languages or varieties the intended potential error is associated. These attitudes and associations may well be different for the various major accent groups. As Figure 3.19 demonstrates, most comments were fairly neutral. Instead of dismissing the errors as unimportant, or describing their negative affect, the “neutral comments” generally either referred to the words or segments

containing the potential errors, or described the misunderstandings that could ensue as a result.

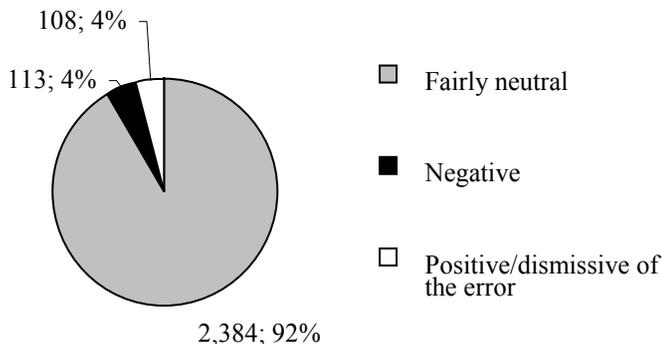


Figure 3.19. Respondents' relevant and possibly relevant comments about individual tokens, divided into three categories according to tone ($n = 2,605$).

While no single token received significantly more positive or negative comments in any major accent group, it is perhaps of anecdotal interest to note that, in keeping with the stereotype of Canadian politeness, the CDN respondents did not volunteer a single negative comment. In addition, the attitude of the IRL judges is also worthy of note. If the incidence of negative comments in the IRL group (19 instead of the expected 7.2) is compared with that of all other groups combined, the difference is statistically significant ($\chi^2 = 21.868$, $df = 1$, $p < .00$). This could conceivably suggest that these judges are either more critical as a group, or feel more prepared in a linguistic experiment to indicate their attitudes to certain pronunciation errors.

Figure 3.20 reveals that most of the comments describing errors in terms of other languages or varieties tend to invoke the context of different accents and dialects of English. For instance, one American respondent came up with the following comments on CAR:

[T]his is acceptable if the speaker has a regional dialect, perhaps Boston. So the question plugs into a whole other area, i.e. attitudes toward stigmatised dialects. I click "relatively unimportant" because I don't want to seem prejudiced, but the fact is that native speakers of Boston come under tremendous social pressure to change their accent (Subject 575).

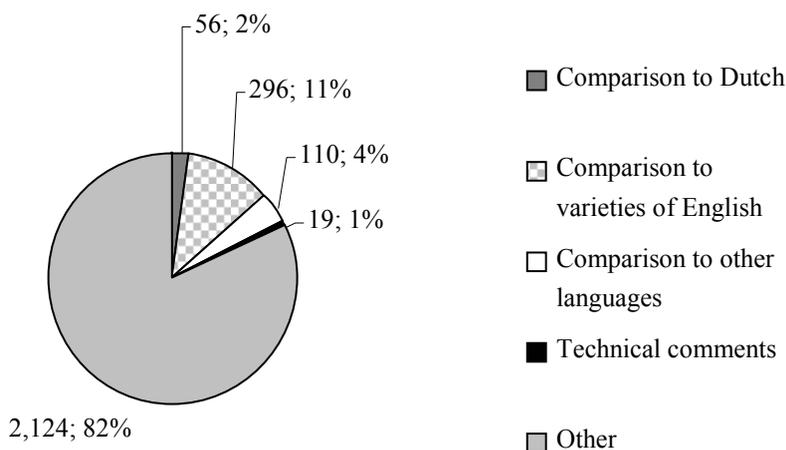


Figure 3.20. Respondents' relevant and possibly relevant comments about individual tokens, arranged in five categories according to content (n = 2,605).

This comment would suggest an interconnection between the affect generated by a token and its occurrence in a particular variety of English – a factor which may cause it to be assessed either more or less leniently by different groups of respondents. If a pronunciation feature is associated with a stigmatised type of English, judges are more likely to assess it negatively than if the realisation is also common in an accent that enjoys some prestige.

In terms of the subject matter of comments, no significant variation was found between major accent groups. A number of individual tokens, however, were singled out for comparison with different varieties of English (see 3.5 for further discussion). Any technical comments (e.g. about the quality of the recording of particular tokens) that may have influenced respondents' judgements will also be discussed in that section.

3.4.4 Overall effect of detection success on severity

Listeners' actual detection of potential errors is another factor which may account for certain variations in the assessment between major accent groups. If respondents detect certain errors more frequently in one version of the experiment than the other, this will affect the severity estimates for that group. These "hit rates" (HR) may be subject to indexical variation (age, sex, leniency, version). Such effects can be estimated by means of multi-level modelling using the program MLwiN. In addition, it would be useful to know how a token was assessed by the different groups of respondents who had actually reported it. This we have termed "adjusted severity" (AS) – the variable also used in the Dutch version of the experiment. It can be established by using the MLwiN program to calculate the effect on different severity estimates of ignoring the

zero severity scores automatically entered for those judges who did not detect the error. There are significant and striking differences between the HR and the AS estimates for different groups, both individually and for all tokens combined.

It should be remembered, however, that it is difficult to distinguish between errors that were detected and not reported, and errors that were not detected at all. In some cases, listeners may have noticed a deviation from the RP or GA norm, but then decided not to report this because they did not think it constituted a “clearly detectable error” (as stated in the instructions). In other cases, they may not have detected any deviation at all, either because it was insufficiently salient or simply because they did not perceive it as an error. All of these cases have been coded with the severity value of zero (“not or incorrectly detected”), and all are clearly extremely useful indications of a particular error’s lack of significance for the respondents in question – even though they may have been motivated by different causes. To compound matters, these cases of “non-detection” may also be hard to distinguish from those instances (coded 1) where respondents stated there was no “clearly detectable error” but proceeded to identify the phenomenon in question in the “space for extra comments”.

It is useful to know which potential errors were reported by listeners in the form of a “hit” (codes 1 to 5), and which were not reported as detected (“no hit”, coded as zero), especially as this affects the general severity estimates. Consequently, it was decided to calculate the HR and AS estimates separately, and, where applicable, to discuss their effect in this and following chapters. Nevertheless, there may be considerable overlap between the different cases of non-detection and those instances where the deviation was reported as detected but not considered an error, significant or otherwise. In the final analysis, all such considerations are important in establishing a hierarchy of error. This is why, in the native-speaker version of the experiment, general severity estimates comprise both the assessment of the errors actually reported as detected (AS) and the level to which they were detected, or reported as detected, at all (HR). In other words, the term “severity” is generally used here to refer to this composite of detection and non-detection – except where it is specified, as in this and following sections, that AS and HR have been calculated separately.

As Table 3.25 indicates, regression coefficients for the HR and the AS show that the sex difference was significant at $\alpha = .05$ ($Z < -2$). The mean HR is estimated to be lower for female respondents by 0.0114 logit units, while their mean AS was lower by -0.08 scale points. The effects of age on HR and AS are also significant ($Z < -2$, $p < .05$). An increase of one year in the age of a respondent leads to an estimated decrease in both the overall HR and the overall AS of 0.006. This corresponds with the results in 3.3.2.1., which show similar significant effects for overall severity (i.e. the composite of hit rate and adjusted severity) by sex and age. As in 3.1.6, a higher self-identified leniency score (signifying an increase in self-diagnosed strictness) corresponds with an estimated increase of the mean HR by 0.093 ($Z = 4.043$, $p < .05$) and of the mean AS by 0.143 ($Z = 7.124$, $p < 0.5$).

Table 3.25. Overall hit rate and adjusted severity coefficients for all tokens (excluding the distractor), broken down by sex, age and leniency. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

	Sex	Age	Leniency
Hit rate coefficient	-0.114	-0.006	0.093
Standard Error	0.045	0.002	0.023
Wald Z	-2.533	-3.000	4.043

	Sex	Age	Leniency
Adjusted severity	-0.080	-0.006	0.143
Standard Error	0.039	0.001	0.020
Wald Z	-2.054	-4.071	7.124

Table 3.26 shows the HR and AS coefficients for each major accent group. Pairwise comparisons of these showed that, after Bonferroni adjustment for multiple comparisons among $k = 7$ group means, none of the differences between these groups reached significance for HR. However, there were significant differences between the AS estimates for GB/RP and US/GA ($\chi^2 = 18.38$, $df = 1$), GB/RP and US/NGA ($\chi^2 = 14.31$, $df = 1$), and GB/NRP and US/GA ($\chi^2 = 11.1$, $df = 1$), all at $\alpha = .05$. Interestingly, the differences between GB/NRP and US/NGA ($\chi^2 = 7.35$, $df = 1$) and between CDN and GB/RP ($\chi^2 = 8.19$, $df = 1$), while considerable, did not reach a level of significance.

Table 3.26. Overall hit rate and adjusted severity coefficients for all tokens (excluding the distractor), broken down by major accent group.

	GB/RP	GB/NRP	IRL	AU/NZ/SA	US/GA	US/NGA	CDN
Hit rate coefficient	0.418	0.326	0.310	0.198	0.200	0.263	0.197
Standard Error	0.088	0.085	0.112	0.129	0.094	0.094	0.100
Adjusted severity	2.697	2.747	2.824	2.750	2.953	2.930	2.913
Standard Error	0.075	0.075	0.097	0.106	0.076	0.082	0.094

This may be connected to the fact that the GB/RP and GB/NRP groups took part in a different version of the experiment from the US/GA and US/NGA groups. If the same calculations are replicated with “version” as an additional

explanatory variable, however, some of the coefficients would be estimated rather differently, as Table 3.27 demonstrates. In this model, a large amount of variance is now subsumed under “version” instead of under “sex”, which may explain why the AS estimate for sex has become positive instead of negative. While women have a significantly lower HR estimate, their overall AS estimate is actually higher. A similar effect may be observed for those taking part in the GA version of the experiment, i.e. all North Americans. This implies that women and North American respondents tended to report the detection of fewer potential errors in the experiment, but were inclined to judge those they did report more severely. The general effects for age and self-identified leniency, however, remain unaltered.

Table 3.27. Overall hit rate and adjusted severity coefficients for all tokens (excluding the distractor), broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

	Sex	Age	Leniency	Version
Hit rate coefficient	-0.660	-0.006	0.189	-0.079
Standard Error	0.031	0.001	0.008	0.032
Wald Z	-21.290	-6.000	23.625	-2.469

	Sex	Age	Leniency	Version
Adjusted severity	0.135	-0.008	0.594	0.255
Standard Error	0.027	0.001	0.006	0.028
Wald Z	5.000	-8.000	99.000	9.107

If the same model is used to estimate the HR and AS coefficients for individual tokens, a similar pattern emerges. Table 3.28 shows (for all tokens except the distractor) which variables have significantly different HR and AS coefficients: sex (men/women), age (younger/older), leniency (self-identified as strict/self-identified as lenient) or version (GA version/ RP version). In all cases, significantly different AS coefficients are all higher for (1) women, (2) younger respondents, (3) judges who identified themselves as strict, and (4) North Americans. Similarly, most or all significantly different HR coefficients are higher for (1) men, (2) younger respondents, and (3) listeners who identified themselves as strict. However, the picture is more diffuse when it comes to version. Some tokens had been reported significantly more readily by North Americans, while others had been noticed demonstrably more often by participants in the RP version of the experiment.

Table 3.28. Incidence of significantly different hit rate and adjusted severity coefficients for eight paired groups of respondents, broken down by token. Right-pointing arrows denote higher coefficients for the first named of the pair; left-pointing arrows (shaded grey) denote lower coefficients for the same group.

Token	Male/Female		Young/Old		Strict/Lenient		GAv/RPv	
	HR	AS	HR	AS	HR	AS	HR	AS
BED	▶	◀			▶	▶	▶	▶
BAT		◀			▶	▶		▶
VAN		◀			▶	▶	▶	▶
WINE		◀	▶	▶	▶	▶		▶
THIN	◀	◀	▶	▶	▶	▶	◀	▶
AUTHOR		◀	▶	▶	▶	▶		▶
BOTH		◀				▶	▶	▶
OFF					◀	▶	▶	▶
THAT		◀			◀	▶	▶	▶
WEATHER		◀				▶	▶	▶
BREATHE		◀			▶	▶		▶
RED		◀		▶	▶	▶		▶
ICE	▶		▶			▶	◀	▶
TIE		◀			▶	▶	◀	▶
DEAD		◀		▶	▶	▶	▶	▶
FILM		◀		▶	▶	▶		▶
CAR	▶	◀	▶		▶	▶	▶	▶
HOT_TEA	▶					▶		▶
NEW	▶	◀			▶	▶	◀	
IMAGIN		◀		▶	▶	▶		▶
PERFECT		◀			▶	▶		▶
TO_WALES		◀			▶	▶	▶	▶
THAT_THA		◀	▶			▶	◀	
SECONDARY		◀			▶	▶		▶
WOULD_ON		◀	▶		▶	▶	◀	
TELL	▶				◀	▶	▶	▶
COLOUR		◀		▶	▶	▶	◀	
STOOD		◀	▶	▶	▶	▶	◀	
INT1		◀			◀	▶		
INT2		◀	▶		◀	▶		
INT3		◀			◀	▶		

Such differences in hit rates help to explain why, in spite of the fact that all relevant adjusted severity estimates are higher for North Americans, the composite severity of a number of a number tokens is still significantly higher for judges in the RP version (as discussed in 3.2.6): otherwise, the overall composite severity would have been higher for North Americans across the board. It also implies that, if a token was ultimately assessed more strictly in the GA version, this is either because more North American listeners had reported the error, or because such listeners considered the error to be more serious. However, if a token was eventually judged more severely in the RP form, this is only because it had been detected demonstrably more frequently by British, Irish or Antipodean listeners. Strikingly, there are no instances of errors being judged significantly more severely by those judges in the RP version who had reported them. In some cases, however, this could be part of a “floor effect”: if too few North Americans reported a particular error, it would be impossible to establish if they assessed it significantly more strictly. The error in *NEW*, for instance, was detected by just ten North Americans.

Apart from being relevant to the token-by-token analysis of severity in 3.5, the above-mentioned effects on the HR and AS estimates, if found to be more generally applicable, give rise to a number of interesting observations about native speakers’ tolerance and detection of Dutch pronunciation errors in English. An example is the effect of sex on overall severity, on HR and on AS. As reported in 3.1.6, female participants’ overall severity was a little lower than that of male participants, which would imply that they accept more deviation from the native standard than do men. This appears to go against what Labov (2001: 266) terms the “general linguistic conformity of women”: “For stable sociolinguistic variables, women show a lower rate of stigmatised variants and a higher rate of prestige variants than men” – provided one accepts the notion that generalisations about the language use of certain groups extend to the way such groups assess language use in others, in particular that of non-native speakers. Labov’s principle does, however, seem to be consistent with women’s higher AS estimates (at least, in the multi-level model with four variables); the female respondents evaluated the errors they reported significantly more severely than their male counterparts. This does not necessarily mean that in a classroom situation, for instance, female judges of non-native speech are much stricter: the present data suggest, after all, that male listeners actually reported more errors, which compensates for their lower strictness to the extent that their “composite” severity is in fact a little higher. Why male participants in this experiment had significantly higher HR estimates for all tokens (except *THIN*) is more difficult to explain.

One may even speculate on whether certain groups of listeners simply favoured different strategies in carrying out error detection and assessment tasks. Whereas some may decide to report even those errors they did not consider to be overly significant, other groups may have wished to distinguish clearly between more severe errors and those too insignificant to report. Such strategies could, of course, also be indicative of more fundamental attitudes to

Dutch-accented English as either “noticeable but not serious” or “serious only where noticeable”. While there is no hard and fast evidence to establish if any group favoured a particular strategy, this may be an interesting avenue for future research.

The effect of age on HR and AS estimates is similar to that on overall severity as discussed in 3.1.6. The significantly lower hit rates for older respondents may be ascribed, at least partly, to the effects of “presbycusis” or “reduction in auditory sensitivity ... that is the hallmark of the aging auditory system” and possibly of “age differences in cognitive abilities” which “also contribute to impaired spoken language processing in older adults” (Sommers 2005: 469). The relative strictness of younger respondents with regard to reported errors (AS) is more difficult to explain. One option would be to view these judges as less experienced with “language variations” (Ryan 1983: 154) and therefore possibly more intolerant of these – although this is very much a moot point.

In addition, self-identified leniency affects HR and AS estimates no differently from overall severity as discussed in 3.1.6. That is to say that listeners who viewed themselves as less lenient had, generally speaking, correspondingly higher HR and AS estimates, with the interesting exception of a few tokens (OFF, THAT, TELL, INT1, INT2, INT3) which were reported significantly more frequently by those labelling themselves as more lenient. Remarkably, none of these tokens were judged particularly severely in any version of the experiment. It is unclear why self-identified lenient judges should report only those errors more frequently that were considered significantly less serious by those who described themselves as “more strict”.

Finally, the question arises why, on the whole, HR estimates should be significantly higher for participants in the RP version (but, nevertheless, with a great many exceptions), whereas AS estimates for the same group should be lower. Whether or not such discrepancies may be viewed as revealing different underlying attitudes to Dutch-accented English as either “noticeable but not serious” (for RP listeners) or “serious only where noticeable” (for GA listeners) is debatable, if only because of those tokens where hit rates in the GA version are demonstrably lower. In some cases, it is the nature or the presentation of the token that appears to be largely responsible for inter-version variation. Some typically Dutch realisations appear to be more stigmatised in GA than in RP, whereas other pronunciations may be viewed as overt Americanisms in one form of the experiment, but not in the other. In some cases (such as BOTH and TIE), the potential error was presented less saliently in one version of the experiment than the other.

Even if these factors do affect hit rates as well as overall and adjusted severity estimates, the fact remains that no reported error was assessed significantly more severely by listeners in the RP form. This is obviously an indication that North Americans object more to clearly identifiable errors than other groups. This may suggest either a lower tolerance of overt foreignisms, a greater emphasis on linguistic conformity, or even both of these. While some

would deplore this as “ethnocentric” or native speaker-oriented, others may applaud it as a transparent attitude which will clarify matters to non-native learners. In any event, it runs contrary to many Dutch people’s perceptions of inhabitants of the British Isles as being more judgemental about foreign accents than North Americans, as was shown in 1.1.

That such perceptions are not universal is demonstrated by Prator’s (1968: 25) very controversial claim that while some English people have a “deep-seated mistrust” of the foreigner “who presumes to speak English too well”, the “mistrust of French and Americans seems rather to be directed toward the outsider who does not speak French or English well”. According to Prator,

If an Englishman is himself a proud speaker of RP, he may find each encounter with a person who obviously does not speak his language well a pleasantly reassuring reminder of the exclusiveness of his own social group. On the other hand, the American’s greater experience with large numbers of immigrants, whose presence in his country he has felt as an economic threat and a social problem, undoubtedly helps to explain his greater antipathy toward foreign accents (Prator 1968: 25).

Even though the factual basis for these claims may be doubtful and the unmistakable biases in the article (ominously entitled “The British heresy in TESL”) may be unpalatable to many readers, Prator’s much-quoted polemic may serve as a warning to those who believe that Americans are much more tolerant of foreign accents than the British. Arguably, it could even be a reminder that, despite the assurances of the proponents of English as a *Lingua Franca*, both tolerant and intolerant native speakers may be motivated, consciously or subconsciously, by desires other than any urge to accommodate non-native learners. In extreme cases, tolerance of accented speech may even be driven by the wish to exclude rather than integrate non-native learners. Anecdotal evidence for this is provided by one Dutch newspaper correspondent (Steketee 2005: 16), who had been assured by “[e]very Briton” he “had ever met ... that the Dutch spoke the English language perfectly”, but who discovered in the course of his seven-year residence in London that “this is the point of the English language: it’s a game for insiders; it keeps foreigners like us out”. Scheuer (2005: 112) also makes the point that a failure to teach non-native learners authentic L1 pronunciation may in fact pander to the exclusionist tendencies found in some native speakers. In her article, she refers to a number of different discussions of this phenomenon, including the observation by Leather & James (1996: 271) that “the foreign speaker’s pronunciation is apparently expected to reflect his outsider role”.

Be that as it may, there is also evidence from other sources that it is not uncommon for Americans to view accented English unfavourably. For instance, Milroy (1994: 179) refers to “a negative and sometimes demonstrably irrational attitude to languages other than English, and by association to English spoken with a ‘foreign’ accent”. This is quoted by Jenkins (2000: 198) as evidence that “negative attitudes exist towards L2 accents of English among members both of

the general public and the ELT... profession". While Jenkins dismisses the views of such "pronunciation experts" as vitiated by "an interest in preserving the phonological status quo", she cannot account for the attitudes of the American "general public" in the same vein (Jenkins 2000: 198). Whether or not American depreciation of accented English is informed by anxieties about immigration, as Jenkins and Prator suggest (see also Milroy 1994: 192ff), such attitudes cannot simply be declared irrelevant by non-native learners faced with native speakers' irrefutable sociolinguistic dominance. Instead, it would be helpful to Dutch learners of American English – and their teachers – to have a realistic appraisal of how foreign accents are judged in the US. They should also realise, however, that while North Americans may be somewhat stricter, they tend not to detect or report all Dutch pronunciation errors as readily as some other groups do.

3.5 Token-by-token analysis

3.5.1 Assessment of BED

There was a general trend for judges to allocate the error of final fortis/lenis neutralisation in BED to the upper ranges of significant errors in this experiment. For instance, this token appears in the top ten of all three hierarchies of error described in 3.2 (overall, RP and GA). This is in accordance with the strong significance normally assigned to fortis/lenis neutralisation for RP (see Brown 1988: 222, Collins & Mees 2003b: 290, Gussenhoven & Broeders 1997: 16) and GA (Collins & Mees 1993: 125) – even though Dretzke (1985: 203) describes it as an intermediate error and it is unclear whether Jenkins (2000: 159) would classify it as non-permissible "approximation". The error is mentioned in only 32% of the pronunciation manuals surveyed by Wrembel (2005: 428).

At the same time, the evaluation of BED was also subject to significant variation between groups of speakers. As has already been shown in 3.2.3 and 3.3.1, it was evaluated significantly more leniently by (1) judges in the RP form of the experiment and (2) female respondents. While the relative leniency of the former is also reflected in Table 3.29 (at least for GB/RP and GB/NRP), the greater strictness of the IRL respondents is particularly striking, as is their similarity to the US/GA judges. One might even be tempted to infer that US/GA and IRL occupy an intermediate position between the other RP and GA accent groups. Following on from this, it is interesting to note that the differences between IRL and GB/RP and GB/NRP are in fact significant at $\alpha = .05$, as are those between the other RP accent groups on the one hand, and US/NGA and CDN (but *not* US/GA) on the other (see 3.4.1). Remarkably, the IRL severity estimate is also characterised by a significantly lower within-group variance among judges, as compared with the GB/NRP judges ($\chi^2 = 10.72$, $df = 1$, $p < .001$) and the US/GA judges ($\chi^2 = 11.05$, $df = 1$, $p < .001$). This suggests that the IRL respondents are more consistent in this respect.

Table 3.29. Severity estimates for BED, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.990	0.118
GB/NRP	2.893	0.137
IRL	3.692	0.177
AU&NZ&SA	2.699	0.259
US/GA	3.474	0.174
US/NGA	3.762	0.126
CDN	3.692	0.177

An analysis of the HR and AS estimates (see Table 3.30) reveals that, of those who had reported the error, it was the women who had tended to judge it significantly more severely. As with most other tokens, this was also true of the less lenient participants and the North American respondents (see 3.4.4). But as the error in BED had been reported significantly more frequently by men (as well as by stricter judges and North Americans), the female respondents' relative strictness is no longer apparent in the composite severity estimate – which is lower for women than for men. Such different estimates for men and women are hard to explain. There is, at least, no doubt that the error was both reported more frequently and evaluated more strictly by participants from the US and Canada.

Table 3.30. Effects on hit rate and adjusted severity coefficients for BED, broken down by sex, age, leniency and version. Significance is obtained if $| \text{Wald } Z | \geq 2$ (in **bold**).

BED	Sex	Age	Leniency	Version
Hit rate coefficient	-0.605	-0.012	0.679	0.981
Standard Error	-0.257	0.010	0.072	0.295
Wald Z	2.353	-1.182	9.429	3.329

BED	Sex	Age	Leniency	Version
Adjusted severity	0.213	-0.007	0.915	0.788
Standard Error	0.097	0.004	0.023	0.100
Wald Z	2.200	-1.783	40.613	7.960

It is difficult to see why the IRL judges should have more consistently evaluated the error as being somewhat more serious than the other accent groups

in the RP form, and why the US/GA respondents judged it a little less severely than the other North American groups. Similarly, it is hard to account for the inter-version variation itself or the difference between male and female respondents – in terms of HR, AS and composite severity. Unfortunately, the relevant or possibly relevant comments for this token (volunteered by 18% of the judges) do not appear to shed any light on this. Virtually all of these (95 out of 100) describe the potential error in fairly neutral terms. As many as eight judges mentioned that the general context of the carrier sentence made the error less serious; another two stated that it also occurred in native English or in Afrikaans. No indication of any affect was found, except perhaps for one IRL judge, who described it as “comical rather than serious” (Subject 395).

One possible explanation for the slight variations in assessment of BED is that it is rarely found in native varieties of English – an exception being bilinguals who speak either Scots Gaelic or Afrikaans (see 4.2.1). As a result, it may well be associated with foreign-accented English. If BED is indeed perceived as a foreignism, the fact that it is slightly more acceptable in Britain and the Antipodes than in Ireland, the US or Canada could perhaps indicate that, at least in this instance, Irish and North American speakers – especially the US/NGA and CDN groups – attach a marginally greater stigma to such foreign pronunciations. At the same time, however, it is perhaps possible that some North Americans associate this pronunciation not with non-native accents but with the even more heavily stigmatised AAVE (African American Vernacular English; see Wolfram & Schilling-Estes 1998: 171, and also 4.4.1). In any event, all major accent groups concur in assigning BED to the upper ranges of significant errors, on a par with, or just below, such high-ranking stress errors as PERFECT and IMAGIN.

3.5.2 Assessment of BAT

There was no disagreement among major accent groups about the significance of BAT; this had been also the case between respondents taking the RP or the GA versions of the experiment (see 3.2.6), between men and women (see 3.3.2) or between younger and older respondents (see 3.3.3). As can be seen in Table 3.31, the severity estimates were consistently high in all cases.

It is not very surprising that such significance should be ascribed to BAT: all native varieties of English appear to maintain a phonemic contrast between /æ/ and /e/ (cf. Wells 2005: 106), which, according to Brown (1988: 221), has the highest functional load of all pairs of vowels in English. Consequently, BAT is frequently described as a very serious error in pronunciation guides for Dutch learners (Collins & Mees 1993: 57, 126, 2003b: 94, 290, Gussenhoven & Broeders 1997: 16). Dretzke (1985: 203), however, classifies it as an “intermediate error”; if BAT is indeed what Jenkins (2000: 159) describes as a vowel sound with a “consistent” “L2 regional quality”, she would classify as “permissible” and therefore not a priority.

Table 3.31. Severity estimates for BAT, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	3.134	0.120
GB/NRP	2.896	0.134
IRL	3.251	0.282
AU&NZ&SA	2.428	0.297
US/GA	3.015	0.191
US/NGA	3.031	0.176
CDN	3.302	0.225

What is somewhat surprising, however, is the uniformity of judgement among major accent groups. While all native varieties maintain an /æ ~ e/ contrast, some realisations of /æ/ (e.g. Australian, New Zealand, South African) are notably closer to BAT than others. (In fact, its realisation has become so very open in mainstream RP (Collins & Mees 2003b: 93) that many dictionaries now transcribe it as [a]; see Weiner & Upton 2000). In spite of that, no significant differences were found, for instance, between the AU&NZ&SA group and any other single group. It is only in a separate post-hoc test between this Antipodean group and all the other groups combined that a significant difference was found ($\chi^2 = 5.71$, $df = 1$, $p < .016$). (This, however, could also be the result of the Antipodean tendency towards slightly increased error tolerance as described in 3.1.6.) In any event, the AU&NZ&SA severity score for BAT still ranks among the Antipodean top ten of significant errors in this experiment.

The uniformity of judgement is borne out by an analysis of the HR estimates, which do not differ significantly by sex, age or version – only, as would be expected, by leniency. Clearly, BAT had not been not detected demonstrably more readily by any of these groups (see Table 3.32). Admittedly, the error had been judged significantly more severely by those women, strict judges and North Americans who had reported it, but this is a recurrent pattern which is true of almost all tokens (see 3.4.4).

Table 3.32. Effects on hit rate and adjusted severity coefficients for BAT, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

BAT	Sex	Age	Leniency	Version
Hit rate coefficient	0.163	0	0.585	-0.145
Standard Error	0.232	0.009	0.063	0.238
Wald Z	0.703	0	9.286	-0.609

BAT	Sex	Age	Leniency	Version
Adjusted severity	0.367	-0.007	0.920	0.681
Standard Error	0.107	0.004	0.025	0.112
Wald Z	3.442	-1.720	36.851	6.099

Of the 105 respondents providing relevant or possibly relevant comments on this token (approximately 19% of the judges), 13 were also reminded of the close realisation of /æ/ in different varieties of English (as in Antipodean Englishes, or in very old-fashioned RP). Seven others associated it with Dutch, German or other foreign languages. A strong affect was found in only a few cases, in comments such as “AAAAAAAAAAAAAAAAARRRGGGHHHH! Sounds German / like the Royal F[a]mily” (Subject 642) and “ridiculously posh” (Subject 247). One respondent from the Irish Republic stated firmly: “Bats are not bets. This [is] the sort of thing that gets people laughed at” (Subject 987).

3.5.3 Assessment of VAN

While VAN was allocated to the upper ranges of significant errors in both versions of the experiment, as indeed it was in the overall hierarchy of error (see 3.2), judges in the GA version considered it to be significantly more serious than their RP counterparts. (No such variation was associated with age or sex – see 3.3.2 and 3.3.3.) The HR and AS estimates for this token show that, apart from the virtually predictable higher AS estimates for women, stricter judges and North Americans, the latter group had also reported the error significantly more often (see Table 3.33).

It is perfectly understandable that VAN should be considered a significant error. The phoneme contrast /f ~ v/ has a relatively high functional load (Brown 1988: 222) and its conflation by Dutch learners has been identified as a persistent source of confusion (e.g. Collins & Mees 1993: 125, 2003b: 290). Since the contrast is maintained in all the self-identified accents provided by the respondents (with the exception of Scots Gaelic), it is unclear why Dutch English /f ~ v/ substitution should be judged differently by respondents taking part in one of the two versions of the experiment.

At the level of major accent groups (see Table 3.34), the inter-version variation is only reflected in the significant differences between US/NGA on the one hand and the two GB groups on the other. None of the pairwise comparisons between GB/RP, GB/NRP, IRL, US/GA and CDN reached a level of significance. While it is difficult to account for the possibly divergent attitude of the US/NGA judges, it was suggested in 3.4.2 that, at least in some speakers, this could perhaps be ascribed to their perception of VAN as a stigmatised or foreign pronunciation (as with BED, OFF and THAT). This may also be true of North Americans as a group.

Table 3.33. Effects on hit rate and adjusted severity coefficients for VAN, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

VAN	Sex	Age	Leniency	Version
Hit rate coefficient	-0.098	0.004	0.524	1.154
Standard Error	0.244	0.010	0.062	0.297
Wald Z	-0.402	0.400	8.452	3.886

VAN	Sex	Age	Leniency	Version
Adjusted severity	0.382	-0.005	0.926	0.749
Standard Error	0.106	0.004	0.025	0.108
Wald Z	3.609	-1.258	36.970	6.919

Table 3.34. Severity estimates for VAN, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.996	0.127
GB/NRP	2.927	0.147
IRL	3.329	0.264
AU&NZ&SA	1.995	0.301
US/GA	3.467	0.158
US/NGA	3.646	0.151
CDN	3.522	0.191

Interestingly, almost all pairwise comparisons involving the AU&NZ&SA group showed significant differences – with the exception of GB/NRP (and GB/RP at a strict $\alpha = .05$). The lower severity estimate for the Antipodean group (as least in comparison to IRL and the three North American groups) may be somewhat puzzling, especially since phenomena such as /f~v/ confusion are not found in any of the Southern hemisphere Englishes either. It is, however, in keeping with their overall tendency towards increased leniency as described in 3.1.6.

21% of the respondents provided relevant comments on VAN, eight of which explicitly referring to it as a characteristically Dutch pronunciation feature. This was also true of two of the four negative comments about this token: “If anything, this is what tends to be irritating about a Dutch accent. It’s also how a Dutch accent is typically taken off by Brits” (Subject 313) and

“This is the most irritating when working with Dutch people” (Subject 740). If these reactions are anything to go by, it is precisely these attitudes that teachers of English in the Netherlands and Belgium should want to protect their Dutch-speaking students from by teaching them to observe the distinction consistently.

3.5.4 Assessment of WINE

Among the major accent groups, there were no significantly different severity estimates for WINE. This is in keeping with the fact that this token did not discriminate in terms of sex (see 3.3.2), age (see 3.3.3) or version (see 3.2.6). Just as in the hierarchies of error described in 3.2 (overall, RP and GA), all major accent groups allocated WINE to the upper ranges of significant errors.

Table 3.35. Severity estimates for WINE, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.614	0.128
GB/NRP	2.800	0.129
IRL	3.185	0.248
AU&NZ&SA	2.273	0.289
US/GA	2.792	0.193
US/NGA	2.845	0.163
CDN	2.877	0.218

As with most other tokens, the AS estimates were significantly higher for women, stricter judges and North Americans (see Table 3.36). While hit rates were predictably higher for the stricter judges, the same was also true of younger respondents (whose adjusted severity was significantly higher as well). Possible explanations for the effect of age on HR and AS estimates have been discussed in 3.4.4.

The severity uniformly associated with WINE is also reflected in the significance traditionally ascribed to this potential error, not just in Collins & Mees (1993: 126, 2003b: 175, 290), but also, in a non-Dutch context, in Dretzke (1985: 203) and Brown (1988: 222). To some extent, its seriousness is also evident from some of the relevant comments generated by this token (volunteered by 17% of the respondents): as many as nine respondents commented quite unfavourably on WINE, five of which stated explicitly that it reminded them of German or a “stereotyped ‘German’ error” (Subject 299). One respondent even went as far as to state that “it will be presumed that the speaker is German as they cannot get *v* and *w* right – might be serious for a Dutch citizen!” (Subject 278). Significantly, another respondent suggested that it made learners “sound like a German in a Second World War film!!” (Subject 346). Admittedly, however, eight respondents actually took the trouble to point out that the error

was not at all serious, as it was very commonly heard or hardly noticeable. One respondent (who described his own accent as “British, very close to ‘ideal’ RP”) even described the error correctly but went on to say: “But relax. English people don’t speak Dutch at all” (Subject 313). Another respondent stated: “Personally, I find this error to be charming in non-native speakers. I think that it is good for non-native speakers to have some elements like this in their speech – it is part of what makes their speech unique” (Subject 393). One may agree or disagree with the attitude of these respondents, but their comments still help to identify WINE as a foreignism. Such perceptions of this token are reinforced by the fact that /w ~ v/ confusion is extremely rare in native accents (see 4.2.4 and 4.4.4).

Table 3.36. Effects on hit rate and adjusted severity coefficients for WINE, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

WINE	Sex	Age	Leniency	Version
Hit rate coefficient	0.215	-0.016	0.497	0.082
Standard Error	0.223	0.008	0.057	0.231
Wald Z	0.964	-2.000	8.719	0.355

WINE	Sex	Age	Leniency	Version
Adjusted severity	0.413	-0.011	0.884	0.510
Standard Error	0.097	0.004	0.023	0.102
Wald Z	4.243	-2.950	38.033	4.996

3.5.5 Assessment of THIN, AUTHOR and BOTH

Even though THIN, AUTHOR and BOTH all represent TH-stopping (in initial, medial and final position), there was a striking dissimilarity between the almost uniformly strict assessment of THIN and AUTHOR as against the more varied evaluation of BOTH. The latter was evaluated significantly more leniently by the groups taking the RP version of the experiment. Whilst this is in keeping with the results discussed in 3.2 and 3.3, it is difficult to explain why stop realisations of /θ/ should be evaluated differently in initial and medial position as opposed to word-finally – other than by emphasising the salience of word onset in perception. It is also striking that far fewer judges in the RP version than in the GA version had identified BOTH correctly. As was suggested in 3.2.6, this may partly be explained by subtle differences in performance between the two actors in the different forms of the experiment.

As was shown in 3.2, THIN and AUTHOR were assigned to the upper and upper-intermediate ranges in all three error hierarchies (overall, RP and GA), and did not differ significantly from each other in any version. It is only in the RP version that BOTH was evaluated significantly differently from THIN and AUTHOR (Wald $Z = 10.6$ and 9.4 respectively) and ranked lower in the hierarchy of error. A comparison of the two versions by token (3.2.6) showed that while BOTH is evaluated significantly more strictly in the GA version, THIN was assessed slightly less leniently in the RP form of the experiment. The three tokens did not discriminate by sex (see 3.3.2.), but older respondents tended to judge THIN and AUTHOR significantly less strictly (see 3.3.3.). A similar effect for age was not observed for BOTH, possibly owing to a “floor effect”: if the overall severity for BOTH had been higher, it would perhaps have been possible to observe a similarly increased tolerance of this token in older respondents.

The findings in 3.2 and 3.3 are confirmed by a comparison of major accent groups, as found in Table 3.37. While THIN and AUTHOR elicited no significant differences between major accent groups, the estimates for BOTH differed significantly between all North American groups and those taking part in the RP version of the experiment. This means that for these tokens, all inter-group variation may be interpreted as inter-version variation.

Table 3.37. Severity estimates for THIN, AUTHOR and BOTH, broken down by major accent group.

Major accent group	THIN		AUTHOR		BOTH	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
GB/RP	3.453	0.092	3.236	0.090	1.615	0.151
GB/NRP	3.534	0.099	3.318	0.102	1.614	0.159
IRL	3.431	0.209	3.088	0.211	1.615	0.322
AU&NZ&SA	3.172	0.245	3.066	0.257	1.305	0.282
US/GA	3.138	0.176	3.236	0.153	3.017	0.183
US/NGA	3.252	0.143	3.368	0.132	3.260	0.143
CDN	3.285	0.172	3.249	0.188	3.043	0.218

In this context, it may be noted that only 52% of the RP judges had identified the potential error in BOTH correctly, as opposed to 88% of the GA judges. This shows that the significantly lower estimate in the RP version is partly the result of the large number of zero scores automatically entered for respondents who had not detected the intended potential error. If most people do

not detect an error, this is clearly an extremely useful indicator of its relative insignificance (see 3.4.4).

The HR estimate is in fact significantly lower for the RP version of the experiment, as can be seen in Table 3.38. This also applies to the AS estimate, which is significantly higher for those women, stricter judges and North Americans who had actually reported the error.

Table 3.38. Effects on hit rate and adjusted severity coefficients for BOTH, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

BOTH	Sex	Age	Leniency	Version
Hit rate coefficient	0.071	-0.005	0.039	1.878
Standard Error	0.190	0.008	0.043	0.237
Wald Z	0.374	-0.625	0.907	7.924

BOTH	Sex	Age	Leniency	Version
Adjusted severity	0.418	-0.007	0.816	0.880
Standard Error	0.098	0.004	0.025	0.098
Wald Z	4.250	-1.870	32.419	8.958

Clearly, BOTH had been reported significantly more frequently by listeners in the GA form of the experiment. This was not the case with the other two tokens. As a few more RP judges (96%) had reported THIN than GA judges (89%), the HR estimates for this token were only just significantly different for the two versions (see Table 3.39). No such significant differences between HR estimates were attested for AUTHOR. It may not be surprising that THIN and AUTHOR had significantly higher AS estimates for women, stricter judges and North Americans – as this is true of most other tokens – but it is noteworthy that younger listeners had also reported them more frequently and assessed them more severely. (See 3.4.4 for possible reasons for this.) THIN is also the only token in the experiment which had been detected demonstrably more often by women (95%) than by men (91%), although the percentages are both so very high as to make speculation about possible differences rather pointless.

A survey of the available literature (cf. 4.2.5 and 4.4.5) shows that there is nothing to suggest that BOTH is more common in any relevant accent than THIN and AUTHOR, or that it is more stigmatised. If anything, /θ/ appears to be subject to TH-stopping more frequently in initial position than when either medial or final. There is no indication of different importance being attached to any of these three positions in pronunciation guides for Dutch learners: neither Collins & Mees (1993: 21–22, 125, 2003b: 141–143, 291) nor Gussenhoven & Broeders (1997: 16, 141–143) differentiate between initial, medial and final

TH-stopping of /θ/ in terms of error significance; the same holds true for Dretzke (1985: 203), Brown (1988: 22) and Jenkins (2000: 137–138). Incidentally, while Gussenhoven & Broeders (1997: 16) and Collins & Mees (2003b: 291) stress the importance of avoiding this error, as does Dretzke (1985: 203), Brown (1988: 222) assigns a relatively low significance to /θ ~ t/ (4 on a rising scale of 1 to 10), and Jenkins (2000: 159) describes it as “permissible” within the context of her *Lingua Franca Core*, although she rather generously allows that “*at the time of writing*, these sounds are still stigmatised in the L1 communities by speakers of RP, GA, and other more standard L1 varieties” (Jenkins 2000: 138, my italics). Even though Crystal (2001: 57) has also raised questions about the “long-term survival of interdental fricatives in standard English, in a world where there will be five times as many English speakers for whom *th* is a pain as those for whom it is a blessing”, Jenkins’s implicit suggestion that the stigmatisation of /θ/ may be suspended in the near future seems premature.

Table 3.39. Effects on hit rate and adjusted severity coefficients for THIN and AUTHOR, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

THIN	Sex	Age	Leniency	Version
Hit rate coefficient	0.886	-0.038	0.927	-0.645
Standard Error	0.341	0.011	0.098	0.327
Wald Z	2.598	-3.454	9.459	-1.97

THIN	Sex	Age	Leniency	Version
Adjusted severity	0.481	-0.017	0.999	0.329
Standard Error	0.100	0.004	0.024	0.106
Wald Z	4.809	-4.333	41.682	3.102

AUTHOR	Sex	Age	Leniency	Version
Hit rate coefficient	-0.097	-0.027	1.061	-0.115
Standard Error	0.339	0.012	0.112	0.352
Wald Z	-0.286	-2.250	9.473	-0.327

AUTHOR	Sex	Age	Leniency	Version
Adjusted severity	0.377	-0.011	0.939	0.455
Standard Error	0.093	0.004	0.022	0.097
Wald Z	4.044	-2.950	42.896	4.672

THIN drew relevant comments from 19% of respondents. The corresponding figure was 15% for AUTHOR, and a mere 12% for BOTH. This effect was probably partly owing to BOTH being detected much less frequently by the judges in the RP version – it is, after all, impossible to comment on an error which one has not detected. While AUTHOR received rather more negative comments than the other two tokens, virtually all such remarks appeared to concern TH-stopping in general, regardless of the position of the /θ/ in the carrier sentence, and were found for both versions of the experiment. Interestingly, the stigmatisation was at times formidable: comments suggested the error made the speaker sound “childish” (Subject 165), “uneducated” (Subject 393), “stupid/weak” (Subject 568), or as if he had a “speech impediment” (Subject 791). One respondent even said that: “It is only serious because the speaker would probably have the piss taken out of him for pronouncing the ‘th’ wrong” (Subject 811), while another suggested that is especially uneducated native speakers for whom TH-stopping is stigmatised: “Educated English speakers understand that non-English speakers have difficulty with ‘th’” (Subject 902). Whereas some judges suggested that other native speakers (especially the Irish, West Indians, New Yorkers, Minnesotans and French-speaking Canadians were mentioned) also had “problems” or “trouble” with this sound, others pointed out that the existence of the phenomenon in certain varieties of English should make it more acceptable and easier to understand. After pointing out that this also applied to other tokens involving /ð/ or /θ/, one respondent stated that this substitution sounded “odd with this accent, but if the rest of the accent has this feature it wouldn’t be an error” (Subject 331). Another observed that “the problem is if an Irishman said it you’d accept it as the rest of the accent would ‘fit’” (Subject 642). In addition, a few respondents denied the error was at all serious: “What is serious when it’s your 2nd language!!!!” (Subject 823). In spite of this, they had in fact detected the error.

Clearly, there is nothing to be found in dialect descriptions, pronunciation guides or in judges’ comments that would unambiguously explain why the RP judges should have assessed BOTH less leniently than THIN and AUTHOR, or why they detected this potential error less readily. For instance, Marslen-Wilson & Welsh (1978: 59) consider that in error identification tasks involving trisyllabic words, “[d]etection responses should be fastest when the deviation occurs late in the word”. If this is also true of monosyllabic or disyllabic words, one would hence expect the error in BOTH to have been identified more readily than THIN and AUTHOR in all versions of the experiment. It does not explain why BOTH was actually detected less frequently by the judges in the RP version only.

If the relative leniency of the RP judges is connected to their lower error detection scores, this could also suggest that the potential error was presented less saliently in the RP version than in the GA form of the experiment. This may have been caused by prosodic differences between the RP and the GA actor. As Fougeron & Keating (1997: 3738) point out, “unsought variation in prosody is a potential confound both within and across speakers”. While great care was taken to ensure that both actors deviated as little as possible from each other in their

renditions of the carrier sentences, post-hoc investigation of the auditory stimuli revealed that, unlike the GA actor, his RP counterpart had divided the sentence containing BOTH into two intonation groups. While the latter says: “We were both young | and inexperienced. ||”, the GA actor does not pause after young, but intones the sentences as “We were both young and inexperienced. ||”. The RP version gives more emphasis to the first nucleus “young”, which may draw attention away from BOTH. In fact, as many as 41 out of 323 RP judges described the token as a stress or intonation error (13%), rather than as a segmental one; strikingly, this was only true of four GA judges (2%). In addition, 19 RP judges even commented on the peculiar or ambiguous stress or intonation of this token. It would seem that in this carrier sentence, stress and intonation were important sources of distraction for the RP judges, but not for their GA counterparts. This may serve to explain some of the differences in the way BOTH was assessed in the two versions of the experiment.

3.5.6 Assessment of THAT, WEATHER and BREATHE

Although THAT, WEATHER and BREATHE all exemplify TH-stopping in different positions, THAT was assessed much less severely than BREATHE and WEATHER in all versions of the experiment (see 3.2.2, 3.2.4 and 3.2.5). No significant inter-version variation was found in the assessment of BREATHE, but WEATHER was judged so differently in the two versions of the experiment that it ranks as the most significant error of the three in the GA version (see 3.2.5) and as less significant than BREATHE in the RP version (see 3.2.4). While none of these tokens discriminated in terms of sex (see 3.3.2) or age (see 3.3.3), the different patterns of inter-version variation are also reflected in the significant differences between major accent groups. This may be deduced from the estimates in Table 3.40. Although there were no significant differences between major accent groups for BREATHE, the estimates for WEATHER were significantly different for all pairwise comparisons except for (1) accent groups within the same version; and (2) IRL ~ CDN (and possibly IRL ~ US/NGA at a more lenient $\alpha = .10$). For THAT, all pairwise comparisons were also significantly different except for (1) accent groups within the same version, as with WEATHER; and (2) any combinations involving US/GA or IRL. (AU/NZ/SA ~ CDN was only significant at a more lenient $\alpha = .10$.) The lack of significant differences for groups such as IRL may be attributed to their intermediate position between the RP and GA groups (at least for some tokens), or to an increased standard error.

While any inter-group variation for THAT and WEATHER (but not BREATHE) points to inter-version variation, it should be noted that some of the variation in estimates for THAT may be the result of this potential error being detected much less successfully in the RP form (by 32% of respondents) than in the GA version (by 58% of respondents). There is a similar discrepancy for WEATHER (RP 55% and GA 93%), but not for BREATHE (RP 63% and GA 66%). This is also apparent from the inter-version differences between hit rates as presented in Table 3.41: North American HR estimates are significantly higher

for THAT and WEATHER, but not for BREATHE (Wald $Z = 0.6758$). Evidently, BREATHE was more or less equally salient to all.

Table 3.40. Severity estimates for THAT, WEATHER and BREATHE, broken down by major accent group.

Major accent group	THAT		WEATHER		BREATHE	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
GB/RP	0.864	0.127	1.623	0.140	2.201	0.169
GB/NRP	0.963	0.145	1.645	0.153	2.239	0.184
IRL	1.189	0.297	2.295	0.297	2.830	0.317
AU&NZ&SA	0.930	0.272	1.292	0.263	2.142	0.312
US/GA	1.496	0.192	3.272	0.145	2.478	0.227
US/NGA	1.960	0.177	3.284	0.134	2.443	0.216
CDN	2.143	0.275	3.041	0.189	2.851	0.343

Table 3.41. Effects on hit rate and adjusted severity coefficients for THAT, WEATHER and BREATHE, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

THAT	Sex	Age	Leniency	Version
Hit rate coefficient	-0.277	-0.011	-0.189	1.048
Standard Error	0.175	0.007	0.042	0.182
Wald Z	-1.583	-1.571	-4.500	5.758

THAT	Sex	Age	Leniency	Version
Adjusted severity	0.581	-0.004	0.734	0.623
Standard Error	0.129	0.005	0.033	0.129
Wald Z	4.510	-0.730	22.294	4.848

WEATHER	Sex	Age	Leniency	Version
Hit rate coefficient	-0.055	0	0.085	2.406
Standard Error	0.200	0.008	0.044	0.290
Wald Z	-0.275	0	1.932	8.300

WEATHER	Sex	Age	Leniency	Version
Adjusted severity	0.339	-0.005	0.795	0.887
Standard Error	0.097	0.004	0.025	0.096
Wald Z	3.501	-1.427	32.370	9.243

BREATHE	Sex	Age	Leniency	Version
Hit rate coefficient	-0.135	0.011	0.200	0.123
Standard Error	0.174	0.007	0.043	0.182
Wald Z	-0.776	1.571	4.651	0.676

BREATHE	Sex	Age	Leniency	Version
Adjusted severity	0.446	-0.007	0.930	0.749
Standard Error	0.130	0.005	0.030	0.132
Wald Z	3.438	-1.380	31.22	5.682

Other than a surprising tendency for self-identified *lenient* judges to report THAT significantly more frequently than strict judges (a result also found with other tokens that had generally been assigned a low severity), the HR and AS estimates revealed no striking departures from the general pattern outlined in 3.4.4. What this means in terms of inter-version variation is that, while listeners in the GA version detected only two of these tokens more frequently, those North Americans who reported them assessed all three significantly more severely.

Apart from the fact that /ð/ occurs in different contexts in THAT, WEATHER and BREATHE, its substitution by /d/ will have different effects in these words. For instance, *breed* and *breathe* form a minimal pair of intransitive verbs that both more or less fit the general context of “The patient began to breathe more regularly”. The relative plausibility of the substitution in BREATHE (and the potential humour resulting from it) may have caused almost as many RP judges (63%) as judges of the GA version (66%) to describe BREATHE as an error. As one respondent noted, the error was even “[m]ore serious than otherwise due to the silly pun that results” (Subject 757).

In British English, *weather* does not have a counterpart such as *wedder*, but this is not necessarily true for speakers of North American English, where medial /d/ is frequently the result of voiced or flapped /t/. Five respondents actually reported possible confusion between *weather* and *wetter*. It should be noted, however, that the carrier sentence “It’s unusual to have such cold weather in August” would be syntactically ill-formed if *weather* were replaced by *wetter*. In any event, the existence of a minimal pair *weather/wetter* may well have affected the perceptions of American and Canadian respondents.

Whilst minimal pairs may help to explain the different estimates for BREATHE and WEATHER, this does not apply to THAT. Neither in American English nor in British English does the word *that* have a counterpart with initial /d/ – apart from the low-frequency technical term *DAT*. As Brown (1988: 218) points out, such “lexical content words” are “unlikely to be confused” with “grammatical function words ... such as *the, those, they, then, though*”, and, by extension, the high-frequency grammar word *that*. It is partly in view of this that Brown ranks the /ð ~ d/ contrast as relatively unimportant (5 on an increasing scale of 1 to 10) (Brown 1988: 222). For similar reasons, Dretzke (1985: 149, 203) rates the German failure to contrast /ð ~ z/ as less important than /θ ~ s/. Whereas Collins & Mees (2003b: 290–291) describe /ð ~ d/ as being equally significant as /θ ~ t/ but more persistent, Gussenhoven & Broeders (1997: 16) ascribe equal importance to these contrasts, while Jenkins (2000: 159) implies that neither contrast is a priority (see 3.5.5). None of the above make a distinction between /ð/ in initial, medial or final position. In any event, the fact that THAT, as a high-frequency grammar word, cannot easily be mistaken for another item may help to explain why it was generally judged less severely than WEATHER or BREATHE.

If some North American respondents noticed initial TH-stopping in THAT a little more readily than their British counterparts, this is more likely to be due to the prevalence of this phenomenon in non-standard American English. A survey of the available literature shows that initial TH-stopping is considered a social marker in American English (see 4.4.5), and this would suggest that North Americans are more likely to object to a realisation which to them may be stigmatised. It should be noted, however, that any such stigma was more evident from the higher US/NGA and CDN estimates for THAT than from the comments inspired by this token. Only 10% of the North American respondents commented on THAT. None of the comments were negative, and one comment was even dismissive of the error: “People from Brooklyn do the same thing. It’s OK” (Subject 299). In the British Isles, this token is more likely to be considered a marker of regional or ethnic identity. Some British respondents associated the phenomenon with Ireland, but others with the speech of West Indian immigrants. One judge even described this realisation rather tendentiously as an “Irish Jamaican problem” (Subject 642). This was actually the only negative comment inspired by this token; the rest of the 6% of RP judges commenting on this token used fairly neutral terms or said the error was easy to understand, harmless or not very noticeable.

Interestingly, both WEATHER and BREATHE generated quite a few more relevant comments than THAT. As many as 18% of North American respondents commented on WEATHER, and 21% on BREATHE, versus 9% and 20% for the respondents in the RP form. Only one RP judge responded negatively to WEATHER (“a speech defect” – Subject 395) as opposed to three GA judges (“not very good English” – Subject 71; “uneducated” – Subject 393; “annoying and distracting” – Subject 1015). TH-stopping in WEATHER reminded 12 judges (9 of which were North American) of other varieties of English, e.g. from Wisconsin

or Minnesota. BREATHE only incited one fairly negative comment: “There are some accents that use **d** for voiced **th**, and then it isn’t an error. But a dental fricative was used in ‘the’ and the rest of the accent is so RP that it sounds hilarious here” (Subject 331). Three respondents were reminded of other accents of English (e.g. New York, Ireland).

If respondents tended to agree on the severity of BREATHE because of the minimal pair *breathe/breed*, and were inclined to judge THAT, as an irreplaceable grammar word, less severely, some of the variation between the RP and GA versions of the experiment may be explained by the absence of a minimal pair *weather/wetter* in British English. In addition, the frequency of initial TH-stopping in non-standard North American English may have caused THAT to be assessed more strictly by some American and Canadian respondents. To some extent, respondents’ comments appear to suggest that this stigma could also have affected the assessment of WEATHER.

3.5.7 Assessment of OFF

Like VAN, OFF exemplifies /f ~ v/ confusion, but in a context where this phenomenon is extremely unlikely to occur in native English: namely in word-final position before a vowel (see 4.2.6 and 4.4.6). While this may be a good enough reason for OFF to be perceived as a foreignism, it should be noted that this substitution occurs in an unstressed high-frequency preposition (*of*) which is generally pronounced with a weak form (/əv/). In addition, replacing *of* by the lower-frequency preposition *off* results in a sentence which, while ungrammatical, will still be perceived as having the same meaning.² This suggests that the substitution in OFF, while clearly foreign, will draw much less attention than that in VAN. This is in keeping with the much lower importance generally attached to OFF in all versions of the experiment, where it was generally allocated to the lower or lower-intermediate ranges of significant errors (see 3.2). The difference in significance between VAN and OFF is not discussed in Brown (1988), Dretzke (1985), Collins *et al.* (1987), Collins & Mees (1993), Gussenhoven & Broeders (1997) or Jenkins (2000). In one textbook, however, Collins & Mees (2003b: 290) specifically include only initial and medial /f ~ v/ confusion as a persistent and highly significant error.

Although there were no statistically significant differences between male and female, or younger and older, respondents (see 3.3.2 and 3.3.3), those taking the RP form of the experiment judged OFF somewhat less severely than those participating in the GA form (see 3.2.6). While the severity estimates for the different major accent groups appear higher in the GA version than in the RP form, it is only the differences between GB/RP and GB/NRP on the one hand, and GB/NRP and US/NGA on the other, that are actually statistically significant (see Table 3.42).

² Whereas the preposition “of” has a frequency of 29391 per 1 million words, the preposition “off” only has a frequency of 214 per 1 million words (see Leech *et al.* 2001).

Table 3.42. Severity estimates for OFF, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	1.289	0.136
GB/NRP	0.735	0.125
IRL	0.962	0.255
AU&NZ&SA	0.953	0.225
US/GA	1.193	0.167
US/NGA	1.538	0.164
CDN	1.445	0.264

In fact, the different estimates for GB/NRP and the US/NGA are the lowest and highest for this token. While the divergence between GB/NRP and US/NGA is in keeping with the inter-version variation noted above, the difference between versions does not help to explain the disagreement between GB/RP and GB/NRP respondents, who both took part in the same form of the experiment. In addition, it is hard to account for why GB/NRP or US/NGA respondents should have judged this token either relatively leniently or strictly – other than by suggesting that, at least for some US/NGA speakers, OFF may be a somewhat stigmatised foreign pronunciation (see 3.4.2 and 3.5.3). Neither do the relatively few relevant comments (made by only 11% of respondents, by 10% in the RP form and 13% in the GA version) shed any light on this. The token generated no more than four positive and/or dismissive comments, and only a single negative one (“*Of* is extremely common and the “f”, when pronounced, is invariably /v/. Thus it’s poor.” – Subject 951).

The relative difficulty of detecting this error may be illustrated by the large number of irrelevant responses to the carrier sentence: as many as 17 judges commented on other aspects of the sentence, including one listener from the American South, who bizarrely claimed that the “female tone of male speaker could mistakenly impl[y] homosexuality” (Subject 944). At any rate, only 37% of RP respondents reported the potential error, as opposed to 50% of North Americans. As Table 3.43 demonstrates, the HR estimate was in fact significantly lower for the latter, as was, perhaps more predictably, the AS estimate (see 3.4.4). This shows that this group’s slightly higher overall severity estimate for this token is based on higher detection scores as well as on increased adjusted severity. (It is somewhat surprising that the HR estimate for OFF was significantly higher for self-identified lenient listeners, but this pattern has also been attested with other tokens that were generally judged to be less serious, such as THAT, TELL, INT1, INT2 and INT3.) Evidently, this is a token that listeners in the GA form (and, it would seem, speakers of US/NGA in particular), find easier to detect and assess as an error.

Table 3.43. Effects on hit rate and adjusted severity coefficients for OFF, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

OFF	Sex	Age	Leniency	Version
Hit rate coefficient	-0.145	-0.001	-0.116	0.405
Standard Error	0.169	0.007	0.041	0.175
Wald Z	-0.858	-0.143	-2.829	2.314

OFF	Sex	Age	Leniency	Version
Adjusted severity	0.206	-0.004	0.741	0.422
Standard Error	0.122	0.005	0.030	0.123
Wald Z	1.692	-0.835	24.662	3.443

3.5.8 Assessment of RED

While both actors had been instructed to pronounce a uvular-**r** in RED, post-hoc analysis of the stimuli revealed that the actor in the RP version has a uvular trill [ʀ], while the GA actor uses a weakly voiced uvular fricative [ʁ]. In spite of these differences between versions, judges of all ages and both sexes concurred in assigning this error to the upper or upper intermediate ranges of significant errors. Judges in the RP version, however, considered RED to be slightly less severe than their GA counterparts (see 3.2.6). This pattern also emerges from the severity estimates for all major accent groups, as presented in Table 3.44.

While the estimates are almost uniformly high, there is a tendency for the GA groups to assess this token a little less leniently. Yet only four major accent groups deviated from each other in statistically significant ways. Both the GB/NRP and the AU&NZ&SA groups had judged the token significantly less severely than the two US groups (the difference between AU&NZ&SA and US/NGA only reaching significance at a more lenient $\alpha = .10$). As was pointed out in 3.4.2, it cannot be inferred from this that, in a comparison with the two American groups, GB/NRP and AU&NZ&SA had evaluated RED less strictly than the other two groups in the RP form (GB/RP and IRL), whose estimates do not differ significantly from US/GA and US/NGA. Nevertheless, it is tempting to relate the comparatively low GB/NRP estimate to the sporadic incidence of uvular-**r** in some conservative varieties of Northern English (see 4.2.7), while it would be interesting to consider the Antipodean estimate in the light of their overall tendency towards greater leniency (see 3.1.6).

Table 3.44. Severity estimates for RED, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	3.186	0.105
GB/NRP	2.942	0.134
IRL	3.275	0.199
AU&NZ&SA	2.819	0.240
US/GA	3.716	0.159
US/NGA	3.687	0.151
CDN	3.562	0.201

Various textbooks on phonetics and pronunciation also ascribe great significance to RED, including Collins *et al.* (1987: 94) and Collins & Mees (1993: 126, 2003b: 291). Dretzke (1985: 207), whose experiment on native English reactions to German-accented English was actually conducted in north-eastern England, places the error in a high “Dringlichkeitsstufe”. He points out that uvular-*r* would have been assessed even less leniently in other parts of the country, where this realisation does not exist at all: “Bei den Informanten in Nordostengland war zumindest damit zu rechnen, daß sie dem uvularen Frikativ eine gewisse Vertrautheit entgegenbringen” (Dretzke 1985: 136–137). As Gimson & Cruttenden (1994: 286) point out in their advice to foreign learners, “a uvular articulation” of /r/, “though interfering little with intelligibility once the listener has adjusted and though still occurring in the speech of some speakers in north-eastern England, is always perceived as unusual”. Gussenhoven & Broeders (1997) do not mention uvular-*r* as a problem, but while Jenkins (2000) does not specifically mention [ʀ] or [ʁ], she unmistakably only allows “rhotic [ɹ] rather than other varieties of /r/” into her *Lingua Franca Core* (Jenkins 2000: 159).

The severe assessment of RED is also reflected in the large number of relevant comments (101) drawn by this token: 20% of all respondents in the RP version commented on RED, and 16% of all GA judges. Interestingly, this token drew virtually the highest number of negative comments on any token: 11, almost all of which had been volunteered by RP judges (who used terms such as “irritating”, “annoying”, “unnecessary”, “dislikeable” and “ugly”). While one respondent pointed out that “this kind of thing inevitably builds up to produce listener fatigue[,] which can explain why some people switch off when foreigners want to talk to them” (Subject 395), another went as far as to say that it sounded German, “which to an Englishman sounds like a crazy professor” (Subject 642). Others were also unpleasantly reminded of French, German or other “European” accents. Five respondents even mentioned Dutch or Flemish – but not necessarily in negative terms.

On the other hand, eight respondents were dismissive of the error and some pointed out that this kind of /r/, variously termed “uvular”, “guttural” or “rolled”, is also to be heard in north-eastern England, the Border Country, the Midlands or in Scotland – compare Wells (1982: 411), according to whom uvular-**r** “can hardly be regarded as a local accent feature” in Scotland, even though he does say that it is “surprisingly common” there. Nevertheless, the concern with this token as evinced in the comments, 24 of which explicitly compared it to a foreign language, together with the uniformly high severity estimates, still suggests that eliminating this error should be considered a top priority for those Dutch learners who use uvular-**r** in English.

In addition, the error was noticed by almost all respondents (92%). There was no significant inter-version difference between the HR estimates, as is shown in Table 3.45. The HR and AS estimates revealed no other striking departures from the general pattern outlined in 3.4.4. This confirms that RED was almost uniformly regarded as a salient error.

Table 3.45. Effects on hit rate and adjusted severity coefficients for RED, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

RED	Sex	Age	Leniency	Version
Hit rate coefficient	0.325	-0.006	0.709	0.412
Standard Error	0.291	0.011	0.077	0.312
Wald Z	1.117	-0.546	9.208	1.321

RED	Sex	Age	Leniency	Version
Adjusted severity	0.274	-0.010	0.932	0.962
Standard Error	0.097	0.004	0.023	0.101
Wald Z	2.834	-2.617	39.774	9.561

3.5.9 Assessment of ICE

Collins & Mees (2003b: 290) list “/aɪ/ over-long before fortis” as a very significant and persistent error, which will cause “words like *bike*, *might*, *type*, *nice*, *knife*” to be perceived as “*/baɪg, maɪd, taɪb, naɪz, naɪv/” (Collins & Mees 2003b: 111). The error in ICE is also mentioned in other textbooks (Collins *et al.* 1987: 96, Collins & Mees 1993: 130), but it does not appear as a significant error in either Gussenhoven & Broeders (1997) or Dretzke (1985). Jenkins’s *Lingua Franca Core* (2000: 145) “altogether eschews considerations of diphthong quality”, and while she stipulates that “whatever quality is used, the length must be that of a diphthong or long vowel, and the variant must be used consistently”, it is unclear if this means that a consistently over-long /aɪ/ as in ICE is acceptable.

Respondents in this experiment, at any rate, generally did not recognise ICE as a particularly important error, neither in the RP or GA version nor in the experiment as a whole. As was shown in 3.2.6 and 3.4.1, North American as well as older and female respondents considered the error to be even less relevant than other respondents, no doubt partly as a result of these groups' significantly lower HR estimates. This is evident from Table 3.46, which also shows that those few North American respondents who spotted the potential error (a mere 19% of listeners in the GA form, as opposed to 39% in the RP version) ranked the error significantly higher (as did the self-identified stricter judges). In spite of this relatively strict assessment, the fact remains that North American respondents found it particularly hard to detect the error at all. Correspondingly, their composite severity estimate is in fact lower.

Table 3.46. Effects on hit rate and adjusted severity coefficients for ICE, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

ICE	Sex	Age	Leniency	Version
Hit rate coefficient	-0.797	-0.028	-0.03	-0.978
Standard Error	0.192	0.008	0.043	0.207
Wald Z	-4.151	-3.500	-0.698	-4.725

ICE	Sex	Age	Leniency	Version
Adjusted severity	0.224	-0.012	0.767	0.703
Standard Error	0.154	0.007	0.031	0.176
Wald Z	1.456	-1.801	24.927	4.000

In addition, there were no significant differences between major accent groups – apart from the Canadians and the GB/NRP group (see Table 3.47). It was suggested in 3.4.1 that this could possibly be used to point to inter-group variation in North America. This is all the more probable in view of the Canadian tendency to raise the diphthong in *ice* to [əi] (see Wells 1982: 494), as a result of which Canadian respondents may well be more tolerant of other non-standard pronunciations of ICE, or even less likely to detect them than other North Americans. Apart from the fact that there is no hard and fast statistical evidence for this, there are no comments by Canadian respondents to support this either. The fact remains, though, that only 4 out of 40 Canadian respondents detected the error successfully.

Table 3.47. Severity estimates for ICE, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	1.127	0.126
GB/NRP	1.396	0.138
IRL	1.090	0.244
AU&NZ&SA	1.411	0.240
US/GA	0.896	0.145
US/NGA	0.937	0.146
CDN	0.528	0.186

In effect, there were relatively few relevant comments on this token (volunteered by 11% of all RP judges and 15% of all GA judges), none of which were negative. If the comments are anything to go by, some respondents appeared to be hard put to describe the error in question: some referred to the “strange” or “funny” intonation or stress in *ice*, whereas others did in fact mention vowel length or quality. There were in fact two positive or dismissive comments, one of which described Dutch English as a legitimate variety of English: “To English speakers one of the telltale signs of a Dutch accent is the pronunciation of ‘ice’ [as] Dutch ‘ijs’... . No-one should regard this as ... a ‘problem’ but rather as yet another variety of international English” (Subject 999). One may agree or disagree with what could almost be regarded as a “prescriptive” view of error tolerance, but it is perhaps telling that this comment was made with regard to an error that was only detected by relatively few respondents.

Whether or not a Dutch pronunciation of English should be equated with a native accent may be a moot point, but it is clear that this particular deviation is less of a priority for native speakers than may be assumed on the basis of textbooks such as Collins & Mees (2003b). It should be pointed out, however, that the significance of the error lies in a word such as *ice* being misinterpreted as *eyes*. Since the judges in this experiment were presented with the spelling of the word *ice* as they heard the carrier sentence, this may have caused them to be less critical. At the same time, such considerations do not appear to have affected judges’ assessment of other potential errors such as BED, BAT and VAN.

3.5.10 Assessment of TIE

The failure to aspirate English initial stops is described as a significant error in a great many textbooks, including Collins *et al.* (1987: 93), Collins & Mees (1993: 125, 2003b: 291) and Gussenhoven & Broeders (1997: 16).³ Jenkins

³ It may be pointed out, however, that aspiration is only mentioned in 37% of the pronunciation manuals surveyed by Wrembel (2005: 428).

(2000: 159) considers aspiration important enough to be explicitly included in her *Lingua Franca Core*, unlike for example /θ/ and /ð/. In spite of this, there were still considerable differences between the various groups of respondents. As was shown in 3.2.6 and 3.4.1, older and North American respondents judged this token significantly more leniently. A breakdown of the severity estimates by major accent groups (see Table 3.48) shows that, with the exception of AU/NZ/SA versus US/NGA, all accent groups that took part in different versions of the experiment also judged TIE demonstrably differently. There were, however, no significant differences between groups taking part in the same form of the experiment. Even a group such as GB/NRP, where unaspirated /t/ may occur in the speech of some respondents (see 4.2.9), did not deviate significantly from any of the other groups in the RP version.

Table 3.48. Severity estimates for TIE, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.380	0.178
GB/NRP	2.104	0.118
IRL	2.090	0.216
AU&NZ&SA	2.351	0.353
US/GA	0.812	0.146
US/NGA	1.183	0.215
CDN	0.742	0.195

As in the case of ICE, the potential error was detected by considerably fewer North Americans (30%) than by judges in the RP version (75%). As Table 3.49 shows, this is evident from the significantly lower HR coefficients for the GA listeners (and self-identified stricter judges). Similarly to many other tokens, the AS estimates were also demonstrably higher for women, stricter judges and North Americans (see 3.4.4). In other words, while the error was detected more in the RP version, it was assessed more severely by those who reported it in the GA version.

Arguably, some of the inter-version divergence in detection rates may be explained by differences in the performance between the two actors. Post-hoc investigation of the auditory stimuli shows that in the carrier sentence “He always wears a tie in the office”, the RP actor gives extra prominence to *tie* by stressing the error word and pausing markedly before it, whereas the GA actor places the nucleus on the first syllable of *office* instead. Even allowing for the focusing function of nucleus location, i.e. the RP version being a response to “WHAT did he wear?”, as opposed to “WHERE did he wear a tie?” in the GA

form, it is still quite possible that placing the nucleus on *tie*, as in the RP version, has contributed to the salience of the error in question. At the same time, it should be noted that both versions of the auditory stimulus had been checked by a trained phonetician for any signs of aspiration in the word *tie*, which were removed where necessary using the speech manipulation program PRAAT (Boersma & Weenink 2002). As a result of this, the absence of aspiration should have been equally striking in either version.

Table 3.49. Effects on hit rate and adjusted severity coefficients for TIE, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

TIE	Sex	Age	Leniency	Version
Hit rate coefficient	0.082	-0.006	0.310	-1.822
Standard Error	0.186	0.007	0.046	0.196
Wald Z	0.441	-0.857	6.739	-9.296

TIE	Sex	Age	Leniency	Version
Adjusted severity	0.333	-0.008	0.834	0.445
Standard Error	0.127	0.005	0.027	0.162
Wald Z	2.620	-1.432	30.620	2.752

Clearly, the lower North American hit rates for TIE cannot simply be accounted for by the relatively small differences between the auditory stimuli. In spite of the importance that textbooks on phonetics and pronunciation aimed at Dutch learners generally ascribe to initial aspiration in English stops, it would appear that an unaspirated /t/ is not as readily recognised by North Americans as it is by other native speakers of English. In fact, only 7% of respondents in the GA version commented on this token, none of which described it in negative terms. Conversely, 16% of judges in the RP version volunteered comments on this token, one of which was negative (“unaspirated Dutch /t/ sounds quaint”; Subject 642). Seven British respondents (and one American) were reminded of Dutch or other foreign languages, while three respondents from Britain associated this feature either with Scottish English or stated that it was “within limits of British variation” (Subject 802).

The North American reaction to TIE shows not only that this potential error was detected less often, but also that a relatively small group of respondents may describe an error as significant, while most judges do not detect it at all. This would suggest that, within certain accent groups, a pronunciation error may sometimes be assessed according to an idealised norm which does not actually affect the judgements of most speakers. It is debatable whether such

idealised pronunciation norms should be allowed to play a role in assigning a hierarchy of error for Dutch learners of, for instance, General American. In fact, it brings into question the notion that initial aspiration of stops is an important acoustic cue which learners of all varieties of English should concentrate on, even those interested in learning only Jenkins's International English.

3.5.11 Assessment of DEAD

While Gussenhoven & Broeders (1997: 131) warn against "overgeneralisation of preglottalisation" without assigning any particular level of seriousness to this error, Collins & Mees (2003b: 290, 153) describe the use of glottal stops with final lenis stops as one of the "most significant" and "persistent" errors which are associated with "more advanced learners". Although this error is also described in other textbooks (Collins *et al.* 1987: 18, Collins & Mees 1993: 14), no explicit reference is made to its significance, but it is pointed out in Collins & Mees (1993: 14) that it "gives the impression to an American of a glottally reinforced strong consonant". As the error in DEAD is not so much the incorrect preglottalisation of a weak consonant (*[deʔd]) as its replacement by a glottal stop (*[deʔ]), respondents listening to this sentence are even more likely to have heard *debt* instead of a *dead*. This would suggest that they will have judged DEAD on the same strict lines as BED.

Severity estimates show that within the different groups of respondents, the patterns of assessment for DEAD and BED are indeed quite similar. The overall estimates for these tokens are not significantly different (see 3.2.2). The same is true for the GA version (see 3.2.5) – but not for the RP form, where the estimate for BED (3.018) is significantly higher than that for DEAD (2.603) (see 3.2.4). It should be noted, of course, that in spite of these relatively small differences, the estimates for BED and DEAD, both in the RP and in the GA versions, are still quite high, with a tendency for North American judges to assess them even more severely. The two tokens were not judged significantly differently by older respondents, but the slightly more lenient assessment of BED by female judges did not extend to DEAD. This would suggest that most respondents, and especially North Americans, attach great significance to potential errors such as final devoicing – and to anything that has the same effect, such as the glottal replacement of a lenis consonant in DEAD.

North Americans' stricter evaluation of DEAD is evident from their significantly higher HR and AS estimates, as can be seen in Table 3.50. As with many other tokens, the AS estimates are also significantly higher for women, younger respondents and less lenient judges. It is also quite common for the latter group to have significantly higher hit rates (see 3.4.4). Nevertheless, the fact that only 11% of GA participants did not report the error (as opposed to 27% of RP listeners) reinforces the notion that glottal replacement of a lenis consonant is particularly salient to North American listeners.

Table 3.50. Effects on hit rate and adjusted severity coefficients for DEAD, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

DEAD	Sex	Age	Leniency	Version
Hit rate coefficient	-0.209	-0.007	0.352	1.204
Standard Error	0.209	0.008	0.051	0.248
Wald Z	-1.000	-0.875	6.902	4.855

DEAD	Sex	Age	Leniency	Version
Adjusted severity	0.262	-0.010	0.926	0.792
Standard Error	0.108	0.004	0.026	0.109
Wald Z	2.430	-2.380	36.345	7.275

A breakdown of the composite severity estimates for DEAD by major accent group reveals a similar pattern (see Table 3.51). The only significant variation in estimates is between groups taking part in different forms of the experiment – with the exception of most pairwise comparisons involving Irish or GB/NRP judges (see 3.4.1).

Table 3.51. Severity estimates for DEAD, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.385	0.149
GB/NRP	2.879	0.159
IRL	2.888	0.282
AU&NZ&SA	2.241	0.297
US/GA	3.345	0.180
US/NGA	3.480	0.156
CDN	3.575	0.182

Although there is no statistical evidence for all pairwise comparisons, the emerging pattern appears to be that of greater leniency, especially for the GB/RP group and the Antipodeans. A relative strictness in the North American and Irish groups was also found for BED. As was suggested in 3.5.1, this is possibly the result of greater stigmatisation of this token as a foreignism, or because of its association with AAVE. In view of the similarity between the two tokens, this

explanation may also be invoked to account for the greater severity attached to DEAD in North America.

An overview of the 101 relevant comments generated by this token (18% of all judges in the RP version, and 19% in the GA version) suggests that many respondents were indeed reminded of *debt* as opposed to *dead*. Two respondents, one from Britain and one from Ireland, found the error either “absurd” or “annoying”, but all other comments were neutral or dismissive. While three respondents referred to it as a foreignism, one respondent stated that the error was “very common in NY” and also “common among African Americans” (Subject 467).

3.5.12 Assessment of FILM

As in the case of RED, the severity estimates for FILM actually show that native speakers attach equal importance to at least some potential pronunciation errors which, while perfectly intelligible, are heavily stigmatised. As demonstrated in 3.2, FILM is generally allocated to the upper or upper intermediate ranges of significant errors in this experiment. In this respect, there were no differences between respondents taking the RP or the GA versions of the experiment (see 3.2.6), between men and women (see 3.3.2.) or between younger and older respondents (see 3.3.3.). As Table 3.52 reveals, while the potential error had not been detected significantly more readily by any of these groups (except for stricter judges), those women, younger respondents, strict judges and North Americans who had reported the error also evaluated it more strictly – but this follows a pattern also found with other tokens (see 3.4.4.)

Table 3.52. Effects on hit rate and adjusted severity coefficients for FILM, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

FILM	Sex	Age	Leniency	Version
Hit rate coefficient	0.284	0.005	0.546	0.343
Standard Error	0.244	0.010	0.062	0.259
Wald Z	1.164	0.500	8.807	1.324

FILM	Sex	Age	Leniency	Version
Adjusted severity	0.324	-0.011	0.905	0.633
Standard Error	0.107	0.004	0.026	0.111
Wald Z	3.040	-2.617	35.377	5.706

Although the error is not mentioned as significant in any of the other relevant textbooks, strong significance is assigned to it in Collins *et al.* (1987: 31, 94) and Collins & Mees (1993: 35, 127). In Collins & Mees (2003b: 291)

it is labelled a “non-persistent” error, which, in spite of its occurrence “in a few English dialects (e.g. types of Scottish, Irish, Lancashire) ... is completely unacceptable [in other varieties] and sounds comic to the overwhelming majority of native English speakers” (Collins & Mees 2003b: 171). This begs the question of whether this stigmatisation is to be seen as extending to these accents as well, or if schwa epenthesis is only stigmatised because of its association with these accents – apart from the problematical suggestion that learners should be warned not to imitate certain accents (or their characteristic features), because other native speakers (particularly of a standard variety) are biased against them. Such an implication would be less loaded if speakers of, for instance, Irish, Scottish or Lancashire dialects applied this stigmatisation to the local accents themselves.

Both the severity estimates and respondents’ comments show that, at least for IRL respondents, FILM is indeed a stigmatised pronunciation. Table 3.53 shows that the IRL respondents have a considerably lower estimate for this error than any other group. In fact, the only pairwise comparisons that are significantly different are those involving the Irish judges (see also 3.4.1).

Table 3.53. Severity estimates for FILM, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.999	0.123
GB/NRP	2.860	0.141
IRL	1.201	0.271
AU&NZ&SA	3.432	0.226
US/GA	3.110	0.166
US/NGA	3.321	0.147
CDN	2.700	0.253

This does not, however, necessarily mean that respondents from Northern Ireland and the Irish Republic assessed this token much less leniently than other groups. Figure 3.21 reveals that if unsuccessful attempts are excluded, the IRL respondents assess the token considerably more severely. This would suggest that, while less than half of the Irish judges detected the error, those that did appeared to be aware of its stigmatisation, or at least let it influence their judgement. This is also apparent from some of the comments volunteered by nine of the 14 Irish respondents who had detected the error, where the stigmatisation expressed is sometimes formidable (“common ... among southern Irish plebs” – Subject 994).

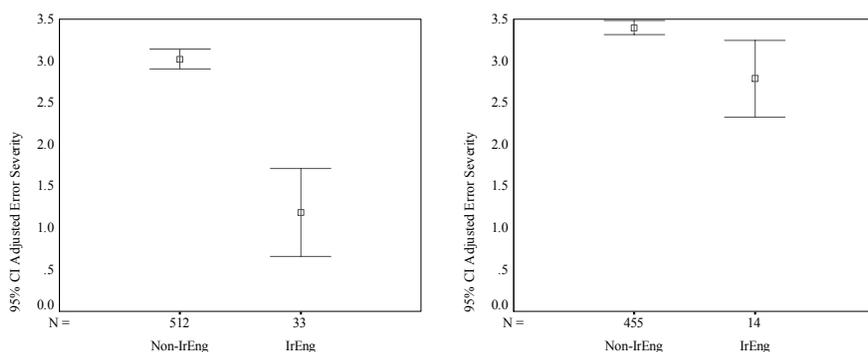


Figure 3.21. Means and error bars (2 standard errors) for the severity scores for FILM, broken down for speakers of Irish English (IrEng) and speakers of other varieties (Non-IrEng). The left-hand diagram includes both successful and unsuccessful attempts to detect this potential error; the right-hand diagram includes the successful attempts only.

Table 3.54. Comments on FILM volunteered by respondents from Northern Ireland and the Irish Republic.

Subject	Accent self-identification	Comment
152	Northwestern Irish	Inserted schwa, acceptable in my own variety of English, some other people say it also, so it [doesn't ??] impede understanding really.
156	Southern Irish	Fil-um
525	British Standard (Irish)	Sounds like "fillin"
657	Southern Irish	Fillum instead of "film", one of the most common errors committed by Irish people.
980	Southern Irish	Fil-lum: common among native speakers too...
987	Irish (south)	The pronunciation filum is normal in Ireland anyway, so does not seem foreign.
994	Southern Irish	But common pronu[n]ciation among southern Irish plebs
995	Southern Irish	The use of l-m in film as separate sounds is common in Ireland but not attractive

The very fact that FILM generated more comments than any other token (31% of all respondents commented on it) suggests that it is very noticeable, and possibly heavily stigmatised. This is also apparent from the negative comments volunteered by other, non-Irish, respondents (see Table 3.55). At the same time, it should be noted that eight judges were actually quite dismissive of the error, some of whom stated that it also occurred in Irish and Scottish English, or

in local dialects. This was actually mentioned in as many as 66 comments, whether or not these were negative, dismissive or fairly neutral.

Table 3.55. Negative comments on FILM volunteered by respondents other than from Northern Ireland and the Irish Republic.

Subject	Accent self-identification	Comment
159	American - West Coast	I actually hear this here, and it bugs me to no end. There should be no shadow vowel after the l in “film”.
160	American - Standard	Film is pronounced “filum” and sounds uneducated.
320	Standard British, RP	/fil@m/ is a stigmatized dialectal pronunciation. Does that make it more serious or less serious?
396	American Southern	Sounds like “fillum” adding a syllable to the word. This is the mark of an uneducated speaker
464	RP	Has the “fillum” found only in irish! VERY VERY serious
585	Upper Midwest	Filum indicates low level of education.
604	British (RP)	Because I’ve been known to say “fillum” for a joke, and people do.
717	British Rec[eɪ]ved Pronu[n]ciation	Sounded like filum... but I have h[ea]rd the word said this way when English folk are fooling around.
757	more or less Standard American	l is more or less ok, but epenthetic schwa is totally impossible for English (it’s cute in Dutch, however).
763	New Zealand	ughhh! “ilm” is one syllable not two
852	Scottish Gaelic English	Class indicator
1019	British London	There are some English people who [woul]d say filim for film, but usually they are considered Yorkshire bumpkins.

3.5.13 Assessment of CAR

It is clear from various pronunciation handbooks, as well as from the different severity estimates for CAR in this experiment, that the insertion of post-vocalic /r/ in a non-rhotic accent such as RP, or its omission in a rhotic variety such as GA, should be considered less of a priority than observing phoneme contrasts with a high functional load, or avoiding highly stigmatised realisations. Neither Collins & Mees (2003b) nor Collins *et al.* (1987) include r-insertion in their top category of “most significant” (Collins & Mees 2003b: 291) or “crucial” errors (Collins *et al.* 1987: 94); Gussenhoven & Broeders (1997: 17) warn teachers not to “go on about it”, and Gimson & Cruttenden (1994: 284) even advise learners

of RP aiming at minimal general intelligibility to retain post-vocalic /r/.⁴ Similarly, Jenkins (2000: 139) opts for “the GA rhotic variant” in her description of English as an International Language, since this is “simpler for both production ... and for reception”. In their textbook aimed at learners of American English, Collins & Mees (1993) do not mention the problem of Dutch learners omitting post-vocalic /r/ in GA, since, “[a]s in Dutch, American English /r/ is pronounced wherever it occurs in the spelling” (Collins & Mees 1993: 33). This may not be true of speakers of variably rhotic varieties of Dutch, or of learners who have had protracted exposure to non-rhotic varieties of English (including those in North America). All the same, it is clear that respondents generally consider **r**-insertion in a non-rhotic accent, or **r**-deletion in a rhotic one, only to be a secondary issue. This is also evident both from the overall estimate for this token and the estimates for the two versions, none of which rank among the fifteen most significant errors (see 3.2). It is true that judges in the RP version, women and older respondents tend to assess CAR even more leniently (see 3.2.6, 3.3.2 and 3.3.3), but these effects are quite small.

A breakdown of the estimates by major accent group (see Table 3.56) shows that the only significant differences are between groups in different forms of the experiment, notably GB/RP and GB/NRP versus US/GA, and GB/NRP versus US/NGA. If British respondents on the whole tend to be a little less severe than GA speakers, and if non-RP speakers (a number of whom are speakers of rhotic accents) are inclined to be a little more lenient than non-GA speakers (some of whom will presumably be speakers of non-rhotic accents), this suggests that **r**-insertion where the prestige variety is non-rhotic is slightly less serious than **r**-deletion where the prestige variety is rhotic.

Table 3.56. Severity estimates for CAR, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	1.852	0.138
GB/NRP	1.593	0.135
IRL	1.968	0.282
AU&NZ&SA	1.714	0.256
US/GA	2.499	0.166
US/NGA	2.424	0.159
CDN	2.303	0.248

This appears to be connected to the fact that while 76% of North Americans detected the absence of /r/, only 57% of judges in the RP version

⁴ Rhotacism is mentioned in only 41% of the pronunciation textbooks discussed by Wrembel (2005: 428).

noticed its presence. As Table 3.57 demonstrates, the latter group have significantly lower hit rates. While, almost predictably, HR estimates were also significantly higher for younger and stricter respondents, it is striking to see that CAR was one of only five errors to be reported significantly more frequently by male respondents. However, the significantly higher adjusted severity estimates for women, stricter judges and participants in the GA form are more in keeping with the established pattern (see 3.4.4).

Table 3.57. Effects on hit rate and adjusted severity coefficients for CAR, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

CAR	Sex	Age	Leniency	Version
Hit rate coefficient	-0.841	-0.023	0.210	0.999
Standard Error	0.185	0.007	0.044	0.198
Wald Z	-4.546	-3.286	4.773	5.046

CAR	Sex	Age	Leniency	Version
Adjusted severity	0.369	-0.007	0.739	0.476
Standard Error	0.107	0.004	0.024	0.103
Wald Z	3.449	-1.781	30.916	4.604

It may be argued that the greater tendency for North American judges to detect deviation from the rhotic norm is caused by the relative scarceness of non-rhotic speakers in North America as a whole, and in this experiment in particular, as opposed to a larger number of rhotic speakers who took part in the RP version of the experiment. Even if that is true, it is not necessary to guess the number of rhotic and non-rhotic speakers in both versions to see that this cannot be the main reason for these higher detection scores. After all, the estimates of the self-identified RP and GA speakers are also significantly different, while both groups are supposed to be equally homogeneous when it comes to rhotacism. It is therefore tempting to conclude that deviation from the rhotic or non-rhotic norm is considered to be more serious in North America, possibly because an *r*-less pronunciation is associated with stigmatised accents such as Boston and New York. There is some support for this in the comments for this token, which were volunteered by 23% of judges in the GA version (as opposed to 14% in the RP version). While 32 Americans and one Canadian stated that the pronunciation feature was also to be heard in other varieties of English, as many as 26 of these explicitly referred to the accents of the Northeast, including Boston, New York and New Jersey, and not always in positive terms (“*car* sounded ‘Brooklyn-ish’ to me and I abhor that accent” – Subject 944; “sounds like a New England accent: very unattractive” – Subject 964). To counterbalance these negative comments, however, there were 10 comments which were

dismissive of this error. As Subject 596 points out, “When I first started hearing Dutch speakers this was funny to me, but I don’t really even notice it any more”.

3.5.14 Assessment of HOT_TEA

Most textbooks describe HOT_TEA as an error of only limited urgency. Collins *et al.* (1987: 93) refer to “[r]eduction of doubled stops” in “hot tea” as a “serious” but not “crucial” error, and Collins & Mees (1993: 24–35) represent it as a “significant” error which is not “of the greatest importance”. In addition, Collins & Mees (2003b: 218) do not include it in their hierarchy of error but nevertheless describe it as a “significant problem”. Furthermore, Gussenhoven & Broeders (1997: 170) also refer to the “[u]ndesirable degemination of double consonants” but exclude it from their “Hints for the future teacher” (Gussenhoven & Broeders 1997: 16–17). Interestingly, Jenkins (2000: 159) actually appears to be highlighting the importance of this type of error by excluding it from her *Lingua Franca Core*, which generally only permits simplification of “medial clusters ... according to L1 rules of elision”.

Respondents in this experiment also consigned HOT_TEA to the intermediate or lower intermediate ranges of significant error. Admittedly, though, these low estimates are largely the result of a general failure to detect the error, rather than it being ranked as insignificant. As many as 64% of respondents did not report the intended error, but those who did assigned it a much higher severity score than is evident from the overall severity estimate. For instance, the AS estimate in the RP version was 3.281 (s.e. 0.154), as opposed to a composite severity of a mere 1.400 (s.e. 0.108), whereas the AS estimate in the GA form was 3.591 (s.e. 0.171), as against a composite estimate of only 1.417 (s.e. 0.139). The composite severity estimates do not in fact discriminate by version (see 3.2.6), by age (see 3.3.3.), or by major accent groups (see Table 3.58).

Table 3.58. Severity estimates for HOT_TEA, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	1.482	0.153
GB/NRP	1.308	0.150
IRL	1.589	0.311
AU&NZ&SA	1.309	0.292
US/GA	1.254	0.189
US/NGA	1.574	0.198
CDN	1.505	0.296

Strikingly, there was a small but significant difference between male and female judges (see 3.3.2.) In fact, only 43% of men detected the error, as opposed to the even lower percentage of 30 of the women. These are also the only groups to have significantly different HR estimates, as can be seen in Table 3.59. The AS estimates for these groups show that, if only these relatively few successful attempts to detect the error are considered, the differences between men and women are no longer significant. This would mean that both sexes rate the potential error uniformly highly, but that female judges detected it a little less readily. The effect, however, is quite small, and is difficult to account for. (Table 3.59 also shows significantly higher AS estimates for stricter judges and North American participants, similarly to what has been attested for many other tokens.)

Table 3.59. Effects on hit rate and adjusted severity coefficients for HOT_TEA, broken down by sex, age, leniency and version. Significance is obtained if $| \text{Wald Z} | \geq 2$ (in **bold**).

HOT_TEA	Sex	Age	Leniency	Version
Hit rate coefficient	-0.669	0.005	-0.031	-0.225
Standard Error	0.175	0.007	0.041	0.182
Wald Z	-3.823	0.714	-0.756	-1.236

HOT_TEA	Sex	Age	Leniency	Version
Adjusted severity	0.301	0.001	0.854	0.711
Standard Error	0.166	0.006	0.035	0.168
Wald Z	1.817	0.177	24.133	4.234

For such a relatively unimportant error, HOT_TEA generated a surprisingly large number of comments (124), with only FILM, PERFECT and TO_WALES ranking directly above it (see 3.4.3). None of the comments were actually negative; most of these were simply attempts to describe a somewhat elusive phenomenon that the format of the experiment made it more difficult to pinpoint as an error. This may explain why so many of those who had detected the error chose to comment on it. There was one only positive/dismissive comment: “Somewhat unusual to an English-speaking ear (erm, if you see what I mean), but not stric[t]ly an error” (Subject 657). While two American respondents identified HOT_TEA as an error that would confuse native speakers, two British judges were reminded of South African English, whereas one judge from the American South (Texas) stated: “The final **t** of ‘hot’ is not distinguished from the initial **t** of ‘tea’. This would be understood by Southern American speakers, but would be considered incorrect pronunciation everywhere” (Subject 826).

Finally, it should be noted that 28% of male respondents commented on this token, as opposed to only 17% of female respondents. This difference between the sexes follows the same pattern as that of error detection. Since none of the comments were negative, it is impossible to establish if the token was possibly more stigmatised by one of the groups. While the variation between sexes cannot be accounted for, the results clearly indicate that HOT_TEA, though significant, is not among the most crucial errors for Dutch learners of any variety of English, and was detected by relatively few native speakers.

3.5.15 Assessment of INDIA

As may be expected, the distractor INDIA was ranked as one of the least significant errors in both versions of the experiment. It would not be appropriate to discuss respondents' attitude to the distractor in detail, or to calculate and discuss hit rates and adjusted severity. Nevertheless, it is noteworthy that there were still significant differences between the two versions (see 3.2.6), notably between GB/RP and US/NGA (see Table 3.60). However, the token did not reflect differences due to sex or age (see 3.3.1).

Table 3.60. Severity estimates for INDIA, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	0.588	0.045
GB/NRP	0.642	0.047
IRL	0.572	0.0908
AU&NZ&SA	0.690	0.083
US/GA	0.737	0.054
US/NGA	0.817	0.045
CDN	0.723	0.075

Such inter-version variation may be accounted for by the tendency for some judges to identify incorrectly an unintended "error". For instance, 17% of respondents (all but one of whom took part in the RP form) objected to the pronunciation of the word "India", while 11% thought there was a problem with either stress or intonation. For these judges, a severity score of "0" was returned – whereas respondents who had detected the distractor correctly had assigned it a "1" (= no error). As a result, the former group would be expected to have a lower severity score than the latter group, something which can indeed be observed in the RP version.

Analysis of the comments show that a minority of respondents in the RP version felt that the /d/ in "India" had not been pronounced clearly, or that it was absent. It is unclear if this feeling was motivated by the lack of a more obvious error, but since no North American judges objected to the /d/ in "India", one

may assume that in this case the inter-version variation may be derived in part from a difference in performance between the two actors.

The lack of any obvious error may have prompted some respondents to comment on the experiment as a whole. For instance, one female self-identified speaker of GA inquired “as to why all sentences are read in a masculine voice – it seems quite sexist” (Subject 695) – from which it is apparent that respondents are not necessarily familiar with, or concerned about, the constraints of experiment design. Another respondent in the same category said she wished “to mention so far that much of the delivery is very nasal” (Subject 588). The “error” of nasality, which was also detected by other respondents with regard to different tokens, is a good example of an American “folk” linguistic attitude to pronunciation. After all, as is pointed out by Collins & Mees (1993: 96), “[a]ll educated American has this slightly nasal quality, and if it’s lacking, people regard this as unpleasant. (Curiously, Americans often term non-nasal speech ‘nasal’, or ‘talking through your nose.’)”.

3.5.16 Assessment of NEW

While in the RP version of this experiment, the intended error was a more characteristically American pronunciation of [nju:] as [nu:], its counterpart in the GA form was the supposedly more British version [nju:]. In this particular instance of token *mirroring* (see 2.1.3, 2.3 and 3.2.6), there is, however, considerable imbalance between the two versions.

Only a small minority of judges choosing the RP version will also say [nu:], for example some speakers from East Anglia, London and New Zealand (see 4.2.14). Conversely, [nju:] is heard very commonly in both the US and Canada (see Wells 1982: 247, 496, and also 4.4.14). According to a pronunciation preference poll quoted in *LPD* (Wells 2000: 510–511), 14% of American judges actually preferred [nju:] to [nu:]. The accompanying graph (Figure 3.22) shows this effect to be even stronger for older Americans. Based on a 1980 survey held by Murray, Lippi-Green (1997: 36) even describes American yod-insertion as a prestige feature, which has “more social currency in the south than it does in the north”. Whether or not this is true, North American respondents will be very unlikely to view [nju:] as an error or a highly stigmatised pronunciation. This may affect their detection and/or severity scores.

Unlike their North American counterparts, judges choosing the RP version are far more likely to view [nu:] as a deviation from the RP norm, and judge the token more severely. Since [nu:] does not impede comprehension and is commonly heard in the English-speaking world, they will not assign a top priority to this token. Understandably, it is therefore not discussed as a significant pronunciation problem in any textbook aimed at Dutch learners of English.

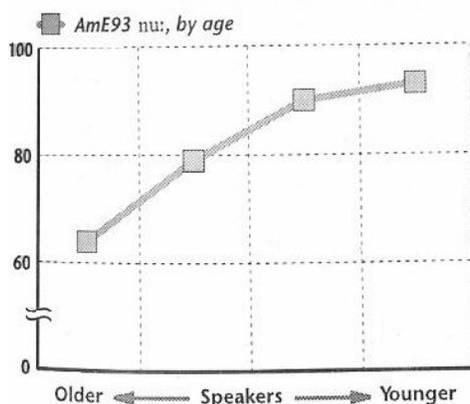


Figure 3.22. Percentage of American judges preferring [nu:] to [nju:], broken down by age group, as according to a poll held in 1993 by Yuko Shitara (reprinted from Wells 2000: 511; permission has been received from the author, 20 May 2005).

Accordingly, relatively few respondents (in either the GA or the RP version) detected an error in this token. In addition, those who reported an error did not generally rank it as very serious. This is evident from the low severity estimates for NEW, which, at least in the GA version, rank among the lowest in this experiment (see 3.2.5). In fact, the difference between the estimates for the RP version (1.457) and the GA form (0.205) is statistically significant (see 3.2.6). The slight overrepresentation of women in the GA version may be why female respondents also judged this token significantly less severely (see 3.3.2).

A breakdown of the estimates by major accent groups reveals that all inter-version variation is significantly different, with the exception of any pairwise comparisons involving Irish respondents (see Table 3.61). The implications of this were discussed in 3.4.1.

Table 3.61. Severity estimates for NEW, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	1.586	0.118
GB/NRP	1.367	0.138
IRL	0.862	0.238
AU&NZ&SA	1.706	0.244
US/GA	0.245	0.059
US/NGA	0.202	0.051
CDN	0.163	0.057

In addition, note that only nine Americans and one Canadian had marked the presence of /j/ as an error, as opposed to the 164 respondents in the RP version who objected to its absence. While inter-version differences in hit rate were in fact strongly significant, nevertheless these did not significantly affect the adjusted severity estimate (see Table 3.62). In other words, North Americans detected the intended “error” much less readily, but it is impossible to establish if those few respondents who detected it also attached less significance to it. As with many other tokens, there are also significant effects of sex and leniency (but not of age) on the hit rate and the adjusted severity (see 3.4.4).

Table 3.62. Effects on hit rate and adjusted severity coefficients for NEW, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

NEW	Sex	Age	Leniency	Version
Hit rate coefficient	-0.795	0.012	0.152	-3.199
Standard Error	0.209	0.008	0.046	0.349
Wald Z	-3.804	1.500	3.304	-9.166

NEW	Sex	Age	Leniency	Version
Adjusted severity	0.422	0.005	0.740	-0.591
Standard Error	0.153	0.006	0.029	0.331
Wald Z	2.750	0.867	25.800	-1.784

While for the vast majority of North Americans, the token went unnoticed, 51% of judges in the RP version detected it, including 12 out of 20 New Zealanders. The comments show that at least 35 British and Irish respondents, as opposed to only two Americans, associated it with regional or social variation. Examples mentioned included Norwich, Norfolk, Lincolnshire, London, New York, the US, Canada and North America. These associations prompted five judges to be dismissive of the error (as were two other respondents), and four to respond to it very negatively (as did one American). This may be seen in Table 3.63, which shows that for at least two respondents, the token is regarded as more of an icon of American speech influence on British English than an example of regional differences in the UK.

Clearly, while North Americans generally do not really notice the difference between the presence or absence of yod in this context, this is not true of British and Irish native speakers of English. It would therefore not be advisable for learners of British English to adopt “yod-less” pronunciations unless their target accent is a regional variety in which this phenomenon is not stigmatised.

Table 3.63. Negative comments on NEW.

Subject	Accent self-identification	Comment
531	British – Southern	This pronunciation is poor English although many lazy English speakers would sound similar, but it is incorrect and sounds bad.
657	Southern Irish	Noo instead of "Nyoo", an error committed by many native English speakers, esp. from the US.
696	US West Coast	This is another sentence in which the speaker seems to come from two different regions of the U.S. so he sounds weird but not incorrect.
879	British-Southern	Mmmm. Bit of an Americanism here. "New" should not be "Noooo". Americans also seem to think the London Underground can be shortened to the "toooooob".
966	Northern British	Normally quite tolerant of spoken English (I have learned to be since moving overseas) I do however take exception to Americanisms such as "noo" for "new" in British English.

3.5.17 Assessment of PERFECT and IMAGIN

There would appear to be little doubt that insufficient mastery of English suprasegmental features can be a strong marker of a foreign accent. Moyer (1999: 100) points out that “intonational and stress errors frequently mark the speaker as nonnative, perhaps more often than segmental errors, due to their significance for discourse fluency”. Anderson-Hsieh & Koehler (1988: 590) argue that “prosody may be more critical than segmentals for comprehension, especially at the fast rate”. Daniels (1995: 83) even goes so far as to say the following:

Most segmental errors, though noticeable, do not interfere with communication. The first and alas, often neglected, priority should be to supply learners of English with ten general and powerful stress rules, because it is at the level of word stress that the errors most damaging to comprehensibility occur.

This is in accordance with the claim made in Gimson & Cruttenden (1994: 274) that for “all learners, accentuation must provide the foundation on which any pronunciation course is built”. Unsurprisingly, it is one of the most commonly mentioned areas of pronunciation difficulty discussed in the manuals surveyed by Wrembel (2005: 428). In addition, as Cutler *et al.* (1997: 151) have pointed out, a large number of studies have documented the significance of incorrect stress as an important factor affecting word recognition in English. Nevertheless, some researchers express reservations about teaching word stress. Jenkins

(2000: 150), for instance, calls it “a grey area” which is only “reasonably important to L1 English receivers” and “rarely causes intelligibility problems” in the data she has collected on non-native interaction in English. Although she goes on to say that “word stress rules are so complex to be unteachable”, she cannot avoid recommending “providing learners with a number of general guidelines” in view of “implications for nuclear stress and sound identification” (Jenkins 2000: 150–151).

If the results of the present experiment are anything to go by, native speakers of English do not consider incorrect word stress placement a “grey area”. Far from considering such errors “reasonably important”, they overwhelmingly judged them to be the most significant. In fact, PERFECT and IMAGIN were assigned the highest severity scores in the experiment, as is apparent not only from the overall hierarchy of error, where they are not even significantly different from each other (see 3.2.2). This is also apparent from the error hierarchies for the RP and GA versions, although here the estimates for PERFECT and IMAGIN are not always significantly different from those for tokens such as THIN in the RP form and TO_WALES in the GA version, which ranked immediately below them (see 3.2.4. and 3.2.5). In addition, the assessment of these tokens appeared to be equally strong for virtually all groups of respondents. For instance, neither PERFECT nor IMAGIN discriminated by major accent group (see Table 3.64). In addition, there was no significant variation for sex (see 3.3.2). Unlike PERFECT, IMAGIN was judged a little less strictly in the GA version (see 3.2.6) and by older respondents (see 3.3.3).

Table 3.64. Severity estimates for PERFECT and IMAGIN, broken down by major accent group.

Major accent group	PERFECT		IMAGIN	
	Estimate	Standard Error	Estimate	Standard Error
GB/RP	3.758	0.080	3.537	0.088
GB/NRP	3.776	0.088	3.619	0.080
IRL	3.860	0.130	3.558	0.153
AU&NZ&SA	3.763	0.174	3.721	0.179
US/GA	3.840	0.101	3.741	0.131
US/NGA	3.865	0.092	3.822	0.109
CDN	3.957	0.135	3.913	0.128

The extremely strict assessment of PERFECT and IMAGIN is in keeping with the importance ascribed to stress in textbooks aimed at Dutch learners (for instance Collins *et al.* 1987: 87, Collins & Mees 1993: 118, 2003b: 291,

Gussenhoven & Broeders 1997: 16). Koster & Koet (1993: 79) also note that both Dutch and English judges of Dutch English object “very strongly to incorrect placement of stress in words”. However, they also state in their conclusion that “current practice in teaching English pronunciation in Holland is correct in paying hardly any attention to suprasegmental features” (Koster & Koet 1993: 90). Similarly, Dretzke (1985: 207) accords incorrect stress placement only an intermediate “Dringlichkeitsstufe”.

Stress errors are clearly very salient to native speakers: the error in IMAGIN was detected by 98% of respondents, while PERFECT was reported by no fewer than 99%. Apart from leniency, no other factors noticeably affected the hit rates for this token. Conversely, the AS estimates were significantly higher for the usual groups of women, stricter judges and North Americans. In the case of IMAGIN, this was also true of younger respondents (see Table 3.65). Evidently, as would be expected, these stress errors were detected almost equally consistently by all groups of respondents, and were assessed particularly strictly by those groups of respondents who tended also to judge other tokens they detected more strictly.

Table 3.65. Hit rate and adjusted severity coefficients for PERFECT and IMAGIN, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald Z}| \geq 2$ (in **bold**).

PERFECT	Sex	Age	Leniency	Version
Hit rate coefficient	-0.267	0.002	1.752	1.472
Standard Error	0.727	0.030	0.279	1.097
Wald Z	-0.367	0.067	6.280	1.342

IMAGIN	Sex	Age	Leniency	Version
Hit rate coefficient	0.839	-0.020	1.248	0.130
Standard Error	0.541	0.018	0.156	0.536
Wald Z	1.551	-1.111	8.000	0.243

PERFECT	Sex	Age	Leniency	Version
Adjusted severity	0.471	-0.003	1.039	0.497
Standard Error	0.101	0.004	0.024	0.105
Wald Z	4.669	-0.875	43.094	4.744

IMAGIN	Sex	Age	Leniency	Version
Adjusted severity	0.436	-0.010	1.005	0.666
Standard Error	0.099	0.004	0.024	0.103
Wald Z	4.417	-2.622	42.621	6.485

The two stress errors also drew a great many comments: 154 for PERFECT and 121 for IMAGIN. These were all relevant, and represented 28% and 22% of all respondents. In fact, both tokens ranked among the top five of most commented on errors. They were a few observations that were negative (five for IMAGIN and eight for PERFECT) or dismissive (three for IMAGIN and one for PERFECT). Incidentally, the latter, volunteered by a Canadian respondent, was illustrative of the almost euphemistically polite reaction to foreign mispronunciations that one would stereotypically associate with natives of that country: “It may just be the way I am used to hearing it” (Subject 190). After all, no native speaker would ever pronounce the token in the way it was done in the carrier sentence (see 4.2.15 and 4.4.15), so it is clearly not just a matter of being either more or less familiar with it.

The dismissive comments drawn by IMAGIN are stated in Table 3.66. While they reveal interesting attitudes to non-native speech and in one case assigning variation in lexical stress to class differences, these remarks appear to lose some of their impact when compared to the sheer numbers of other native speakers detecting, rejecting and commenting on the errors in question.

Table 3.66. Dismissive comments on IMAGIN.

Subject	Accent self-identification	Comment
313	British, very close to “ideal” RP	Serious only if assessed against native upper-class standards in SE Britain, but can be heard throughout the world of new Englishes. Therefore, I wouldn’t strictly regard it as an error.
696	US West Coast	Any native speaker would understand and, we hope, help the non-native speaker correct this trivial error.
980	Southern Irish	Doesn’t impede comprehension... but clearly marks speaker as foreign.

3.5.18 Assessment of TO_WALES, THAT_THA and WOULD_ON

While Gimson & Cruttenden (1994: 281) stress the importance of “the correct reduction of unaccented grammatical items” as belonging to “the base for the teaching of the language’s pronunciation”, Collins & Mees (1993: 102–103, 2003b: 20) also warn Dutch learners to avoid overusing strong forms, as do Collins *et al.* (1987: 75) and Gussenhoven & Broeders (1997: 16). Weak forms are mentioned in 52% of the pronunciation textbooks discussed by Wrembel (2005: 428). Nevertheless, Koster & Koet (1993: 78) found that neither Dutch nor English judges felt “the absence of weak forms” to be “very annoying”. There are even those who think the significance of weak forms has been overstated. Jenkins (2000: 146), for one, expresses doubts about their pervasiveness in native-speaker English as well as their teachability. While it is hard to verify

such claims (see 4.2.17 for a more detailed discussion), avoidance of weak forms may nevertheless be relatively difficult to diagnose as an error. This is apparent from the striking differences in evaluation of these three tokens.

For instance, while phonemic errors may be obvious from a single occurrence in a carrier sentence, the effect of the lack of weak forms could well get to be more salient through regular repetition. Such long-term effects cannot easily be established within the framework of the present experiment. Similarly, there are cases in which the use of strong forms, instead of being perceived as an error, actually changes the meaning of a sentence. In those cases, a disambiguating context could be added to rule out any such options. An example is the addition “for a long relaxing holiday” to “I want to go /tu:/ Wales”. Arguably, this added context effectively excludes the possibility of contrastive stress for *to* as opposed to *from* – although one listener from the Midwest still felt that “[t]his sentence stress and intonation could be correct in response to a question like “Are you going *from* Wales?” (Subject 276, italics added). If the design of the experiment had allowed it, a longer or more elaborate context could have been supplied for additional clarification.⁵

A further problem is that, when asked, linguistically naive native speakers may not trust their intuitions and proscribe the use of weak forms as a deviation from what they perceive as the norm. This is a well-attested phenomenon (Gimson & Cruttenden 1994: 281, O’Connor 1971: 117). To take an example from popular fiction, the writer J.K. Rowling (1999) indicates that one of the characters in the *Harry Potter* series is an uneducated rough diamond through presenting his speech as a mixture of weak forms and sub-standard English. Needless to say, it would in fact be unusual for any native speaker to pronounce “for” in any other way than /fə/ or /fɜ:/ – presumably what Rowling implies by the eye-dialect representation “fer”.

“School gov’ners have bin told, o’ course,” said Hagrid miserably. “They reckon I started too big. Shoulda left Hippogriffs fer later... done Flobberworms or summat... jus’ thought it’d make a good firs’ lesson... s’all my fault...” (Rowling 1999: 92).

Since such perceptions of weak forms are by no means uncommon in linguistically naive judges, Gimson & Cruttenden (1994: 281) point out that, in this respect, “it is wiser to listen to the way in which the native speaks rather than

⁵ As one participant put it, “all of the sentences used can’t be judged well, because we have no context to put them in. This is a huge methodologi[c]al problem” (Subject 962). It is true that, in the case of suprasegmental phenomena, a more elaborate general context could possibly have helped respondents to evaluate the severity of such errors. It would, however, be much more difficult to envisage a context that would affect judges’ evaluations of segmental errors in a similar way. In any event, the addition of these would have made the experiment considerably longer and therefore likely to be much less attractive to respondents.

ask his opinion". Actually, judges may even perceive strong forms merely as stylistically awkward or hypercorrect rather than as a serious error, as Dretzke (1985: 172) found in his experiment on native-speaker judgements of German-accented English.

Finally, the use of strong forms cannot easily be separated from other phenomena such as vowel gradation and connected speech stress, and may therefore have different effects in the three carrier sentences. For instance, the use of the strong form /tu:/ in "I want to go to Wales for a long relaxing holiday" is more likely to be perceived as an error in sentence stress than is true of the substitution */ðæt ðæt/ for /ðət ðət/ in "They all said that that may be done very differently". While the use of /tu:/ may be interpreted as the incorrect use of emphatic or contrastive stress, this is unlikely to apply in the case of a complementiser such as "that". Similarly, the marked use of the strong form /wɒd/ in "I'd like to tell her what he's up to, but she would only go and let the cat out of the bag" may cause this word to be perceived as carrying emphatic stress. This makes it more difficult separately to observe the different effects of stressing and avoidance of weak forms. In addition, post-hoc inspection of the auditory stimuli reveals that in the same carrier sentence, the word "only" also attracted some stress – a tendency particularly apparent in the RP version, which may have affected the salience of the intended error. These subtle differences in stress patterning have persisted in spite of efforts to reduce them using the program PRAAT (Boersma & Weenink 2002). The use of speech manipulation in this context was presumably what two respondents were referring to when they observed that this carrier sentence sounded "like a machine" (Subject 217) or, contentiously, "like Ste[ph]en Hawking" (Subject 642). In any event, any such differences will inevitably make a comparison between versions of the experiment less reliable.

These different considerations may help to explain why the tokens illustrating the incorrect use of strong forms were not judged to be particularly severe – except where they could also be interpreted as errors in sentence stress (or at least as marked deviations from expected stress patterns). This is true in the case of TO_WALES and, to a much lesser extent, for WOULD_ON. While TO_WALES ranked as one of the most significant errors in the upper intermediate ranges of the overall hierarchy of error, WOULD_ON and THAT_THA were assigned to the lower intermediate ranges (see 3.2.2.). Similar patterns were found in the RP and GA versions, except that in the former THAT_THA was evaluated as significantly less severe than WOULD_ON. In fact, TO_WALES was judged significantly more strictly in the GA version. While it ranked among the top five of significant errors in the RP version (see 3.2.4), it was among the top three in the GA form (without being significantly different from the errors ranked directly above or below; see 3.2.5). At the same time, WOULD_ON and THAT_THA were evaluated more severely in the RP form (see 3.2.6). There were no significant effects for sex on the composite severity estimate (see 3.3.2), but older respondents assessed WOULD_ON and THAT_THA demonstrably more leniently than younger ones (see 3.3.3).

A breakdown of composite severity estimates by major accent group (see Table 3.67) reveals that in the case of TO_WALES, it is only those for US/GA and US/NGA that are considerably and significantly higher than those for GB/RP and GB/NRP. For THAT_THA, it is only the difference between the stricter GB/RP speakers and the more lenient US/GA judges that is statistically significant. If this suggests that GB/RP judges tend to evaluate THAT_THA more strictly than other groups in the RP version, it should be remembered, as is pointed out in 3.4.2, that pairwise comparisons between these groups are not statistically significant for these tokens. In the case of WOULD_ON, however, all groups taking part in the GA version assessed this token significantly more leniently than any groups in the RP form. The inter-version variation between these three tokens is illustrated by the significantly different hit rates.

Table 3.67. Severity estimates for TO_WALES, THAT_THA and WOULD_ON, broken down by major accent group.

Major accent group	TO_WALES		THAT_THA		WOULD_ON	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
GB/RP	3.200	0.104	1.654	0.156	2.177	0.142
GB/NRP	3.140	0.104	1.421	0.148	1.882	0.149
IRL	2.769	0.237	1.362	0.270	2.044	0.276
AU&NZ&SA	3.203	0.174	1.523	0.272	1.967	0.254
US/GA	3.681	0.090	0.862	0.163	0.722	0.147
US/NGA	3.679	0.112	1.186	0.165	0.742	0.139
CDN	3.552	0.166	1.237	0.257	0.688	0.189

Both the differences in assessment between tokens, and the inter-version variation for each token, are directly connected to the consistently and significantly different hit rates for each token in the two forms of the experiment. While the error in TO_WALES had been reported by no fewer than 95% of all judges (as many as 99% in the GA version as opposed to 93% in the RP form), WOULD_ON had been only detected by 51% of all respondents (only 26% of GA listeners as against 68% of RP listeners), whereas THAT_THA had only been reported by 45% of all judges (35% in the GA version versus 52% in the RP form). These differences in hit rates are all statistically significant, as can be seen in Table 3.68.

Table 3.68. Hit rate and adjusted severity coefficients for TO_WALES, THAT_THA and WOULD_ON, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald Z}| \geq 2$ (in **bold**).

TO_WALES	Sex	Age	Leniency	Version
Hit rate coefficient	0.515	-0.002	0.800	1.924
Standard Error	0.386	0.015	0.093	0.616
Wald Z	1.334	-0.133	8.602	3.123

TO_WALES	Sex	Age	Leniency	Version
Adjusted severity	0.314	-0.006	0.926	0.730
Standard Error	0.094	0.004	0.023	0.096
Wald Z	3.358	-1.639	40.928	7.596

THAT_THA	Sex	Age	Leniency	Version
Hit rate coefficient	0.005	-0.027	0.022	-0.662
Standard Error	0.172	0.007	0.041	0.180
Wald Z	0.029	-3.857	0.537	-3.678

THAT_THA	Sex	Age	Leniency	Version
Adjusted severity	0.546	-0.008	0.788	0.247
Standard Error	0.136	0.006	0.031	0.152
Wald Z	4.015	-1.403	25.299	1.627

WOULD_ON	Sex	Age	Leniency	Version
Hit rate coefficient	0.007	-0.018	0.245	-1.77
Standard Error	0.185	0.007	0.045	0.197
Wald Z	0.038	-2.571	5.444	-9.000

WOULD_ON	Sex	Age	Leniency	Version
Adjusted severity	0.374	-0.009	0.830	-0.314
Standard Error	0.129	0.005	0.028	0.162
Wald Z	2.898	-1.768	29.692	-1.940

Younger respondents also had significantly higher HR estimates for THAT_THA and WOULD_ON, while stricter judges detected TO_WALES and WOULD_ON demonstrably more often. It is interesting to note that while in the

case of the *TO_WALES*, the AS estimates are significantly higher for the usual groups of women, strict judges and North Americans, this is not true for the other two tokens – at least where the latter group is concerned. For *THAT_THA* and *WOULD_ON*, adjusted severity did not discriminate by version (as was the case for *NEW*, *COLOUR*, *STOOD*, *INT1*, *INT2*, and *INT3*). This means that for these two tokens, the higher composite severity estimate in the RP form is solely the result of higher detection scores rather than of more severe judgements.

The trend among RP listeners towards a stricter evaluation of *THAT_THA* and *WOULD_ON* is illustrated by the percentage of respondents providing relevant comments on these token. While *THAT_THA* prompted 13% of judges in the RP version to comment, as opposed to 10% in the GA version, *WOULD_ON* drew observations from 14% of RP respondents as against 8% in the GA form. However, *TO_WALES* generated many more relevant remarks in both versions (140 in total): 27% in the RP version versus 24% with North American listeners. The latter was in fact among the top three of most commented on tokens – while it also drew a few irrelevant observations, most notably “Sorry, can’t have a relaxing holiday in Wales” (Subject 381) as provided by an Irish respondent. It also generated eight negative remarks, and a single dismissive one. The lower number of relevant comments drawn by of *THAT_THA* (65) and *WOULD_ON* (62) is also in keeping with the patterns noted above. While they also generated a number of irrelevant remarks (particularly *WOULD_ON*), it was especially striking that no fewer than eight North Americans reported TH-stopping in one or both of the two occurrences of “that”, as opposed, or in addition to, the use of a strong form. This was not reported by a single listener in the RP version – either as a result of this actor’s performance or for whatever other reason. For instance, it would be interesting to discover if a speaker’s failure to use weak forms may actually cause listeners to hear a segmental error (especially a highly stigmatised one, as in this case) in addition to, or instead of, a suprasegmental error. This would suggest that while listeners may not detect the absence of weak forms as an error in itself, they may well perceive the effects this has on sentence stress or possibly on segmental features. The data show that this effect is extremely strong in the case of *TO_WALES*. It would therefore be premature to claim that avoidance of weak forms is not a serious source of error.

3.5.19 Assessment of *SECONDAR*

As a rule, textbooks aimed at teaching RP do not describe realisations such as [ˈsekəndəri] as important errors. The same is true of [ˈsekəndrɪ] in GA textbooks. This is hardly surprising, since both are generally recognised to be native-speaker realisations associated with high-prestige varieties of English. Their inclusion in this experiment, however, is warranted by the relative importance attached by some of the Dutch judges to quadrisyllabic pronunciations of words such as *secondary* and *secretary* (see 2.1.3 and 2.3). If this is motivated by objections to the use of General American features in RP-modelled Dutch English, it would be useful to know if such attitudes are shared by different groups of native speakers. The design of this experiment makes it possible to

compare the extent to which judges in the RP and GA versions object to such examples of “dialect mixing”. This has been done by providing a GA realisation of *secondary* in the RP form and an RP realisation in the GA one – a procedure referred to as *mirroring* (see 2.1.3, 2.3 and 3.2.6).

In the overall hierarchy of error, native-speaker judges assigned SECONDAR to the lower-intermediate range of significant errors (see 3.2.2.), with similar rankings for the RP and GA versions (3.2.4. and 3.2.5.). There was no significant difference between versions, which is why it was suggested in 3.2.6. that the RP realisation of *secondary* is as unproblematical to North American listeners as the GA pronunciation is to respondents in the RP form. In addition, the token did not discriminate by sex (see 3.3.2), age (see 3.3.3) or major accent group (see Table 3.69). Evidently, there was no demonstrable difference between groups of native-speaker judges in giving a relatively low priority to this example of dialect mixing.

Table 3.69. Severity estimates for SECONDAR, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.129	0.128
GB/NRP	1.939	0.133
IRL	1.897	0.270
AU&NZ&SA	1.757	0.260
US/GA	1.874	0.171
US/NGA	1.863	0.163
CDN	1.580	0.244

In addition, the potential error in SECONDAR was not reported significantly more frequently by any one group (other than the self-identified stricter judges) – somewhat unusually, there was no effect of version on the HR estimate (see Table 3.70). While the AS estimates were significantly higher for the women, stricter judges and North Americans, this pattern has been found with many other tokens (see 3.4.4). In spite of these predictable differences between AS estimates, it is still striking that the intended error appeared to be equally difficult to detect for listeners in both versions.

The relevant observations generated by SECONDAR do not offer any further insights. While the token drew as many as 100 relevant observations (volunteered by 18% of all respondents), virtually all of these were neutral – some even dismissive. In point of fact, the number of comments was precisely the same in both versions (50 each), albeit that these represented 16% of judges in the RP version and 23 in the GA form. One American judge stated, somewhat

dramatically, that “Americans use clearly British pronunciations at their peril” (Subject 575), but this is exceptional. In addition, some respondents in the GA form associated the RP realisation with other varieties of North American English, but it is unclear if this is based on adequate familiarity with these accents (see 4.4.18). At any rate, this phenomenon was not attested in the RP form.

Table 3.70. Hit rate and adjusted severity coefficients for SECONDAR, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald Z}| \geq 2$ (in **bold**).

SECONDAR	Sex	Age	Leniency	Version
Hit rate coefficient	0.144	-0.004	0.295	-0.277
Standard Error	0.183	0.007	0.046	0.188
Wald Z	0.787	-0.571	6.413	-1.473

SECONDAR	Sex	Age	Leniency	Version
Adjusted severity	0.472	-0.002	0.739	0.322
Standard Error	0.100	0.004	0.023	0.104
Wald Z	4.710	-0.559	31.604	3.088

The lack of significant inter-version differences in composite severity and hit rates, coupled with fairly low composite severity estimates for all groups and the general absence of any emotive comments, suggests that the intended error in SECONDAR should not be given considerable priority in teaching pronunciation to Dutch learners of English. In this respect, it makes no difference if these learners’ model is RP or GA. It may be true, as Van der Haagen (1998: 96) found with a number of secondary school pupils in the Netherlands, that Dutch learners attach more prestige to a four-syllable pronunciation of words ending in *-ary*, even if they use three-syllable realisations more frequently. In this instance, however, it makes little difference to the native speaker which of these pronunciations is adopted.

3.5.20 Assessment of TELL

As Collins & Mees (2003b: 171) point out, general pronunciation textbooks are more concerned with encouraging learners to produce dark [ɫ] rather than with discussing the effect the strongly pharyngealised [ɫʰ] of Dutch English may have on native speakers. This attitude to dark [ɫ] is indeed evident from the “Advice to foreign learners” provided in Gimson & Cruttenden (1994: 185) – although a warning is sounded against over-velarisation in the English of speakers of Slav

languages.⁶ As Collins *et al.* (1987: 30) state: “Dutch speakers use a dark [ɫ] which is much too dark to be acceptable in English; it will usually be understood, but it tends to sound ugly to an English ear” (30). The effect has been facetiously compared to a “whale gargling with treacle” (Beverley Collins, personal communication). While Collins *et al.* (2001: 28) warn students of RP about the “bad effect” the “wrong kind of /l/ can have ... on your English accent”, they also emphasise to learners of American English that the “hollow” quality of Dutch and its “back-vowel effect may give the impression that /l/ is missing altogether” (Collins & Mees 1993: 34).

In spite of this, such effects remained largely unnoticed by the respondents in this experiment – particularly in the RP version, where only seven participants reported the error (2%), as opposed to 54 judges in the GA form (24%). (Interestingly, six of these seven respondents were speakers of RP or other varieties of Southern British English.) It almost goes without saying that such striking differences between versions are statistically significant, as is further demonstrated in Table 3.71. (The error was also reported more frequently by men and by more lenient judges, while those instances that were reported were assessed more strictly by more severe judges and by North Americans.)

Table 3.71. Hit rate and adjusted severity coefficients for TELL, broken down by sex, age, leniency and version. Significance is obtained if $| \text{Wald } Z | \geq 2$ (in **bold**).

TELL	Sex	Age	Leniency	Version
Hit rate coefficient	-1.085	0.016	-0.821	1.638
Standard Error	0.278	0.009	0.082	0.283
Wald Z	-3.903	1.778	-10.012	5.788

TELL	Sex	Age	Leniency	Version
Adjusted severity	0.211	-0.006	0.393	1.730
Standard Error	0.215	0.008	0.066	0.233
Wald Z	0.978	-0.742	5.977	7.399

Such low hit rates have affected this token’s placing in the different hierarchies of error: while its rank is the lowest but one in the overall hierarchy (see 3.2.2), it is the very lowest in the RP version (see 3.2.4) and among the lowest of the intermediate errors in the GA form (see 3.2.5). Whereas the composite severity estimates were significantly different for the two versions (see 3.2.6),

⁶ Dark I was treated in only 18% of the pronunciation materials available in Poland which were surveyed by Wrembel (2005: 428).

there were no effects of sex (3.3.2) and age (3.3.3). In addition, the only significant variation to be found between major accent groups was between those groups taking part in different forms of the experiment; the exception being all but one pairwise comparisons involving Canadian respondents – the difference between GB/RP and CDN was in fact significant (see Table 3.72). There is no statistical evidence to establish clearly if Canadians assessed the token differently from other North Americans, but it may be noted that only four judges from Canada actually reported the error (10%). This does suggest that in this respect (as in many others), they should not be treated on a par with US respondents.

Table 3.72. Severity estimates for TELL, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	0.079	0.041
GB/NRP	0.010	0.021
IRL	0.050	0.064
AU&NZ&SA	0	0
US/GA	0.748	0.142
US/NGA	0.837	0.144
CDN	0.341	0.171

The error in TELL generated only 32 relevant comments (drawn from 6% of all listeners), none of which were negative and one of which was dismissive: “So far these little accent things are co[m]pletely unimportant. People have a million accents and why should a Dutch speaker of English sound like one from Portland, Oregon [?] Wouldn’t I like to know that you’re Dutch?” (Subject 696). Interestingly, this presents accent as a totally unproblematical and handy detection mechanism of foreign speech, but one wonders if this respondent’s reaction would have been the same if he had been confronted with a carrier sentence containing a combination of some of the thirty-one errors featured in the experiment. In any event, three other observations also showed that respondents associated the intended error with Dutch English.

The dramatic disparity in detection and assessment between American listeners and others may be attributed to a number of factors. As was already suggested by Collins & Mees (1993: 34), the effect of Dutch dark [ɫ] may be that of L-dropping or L-vocalisation, which is a stigmatised realisation to be found in a number of US accents (see 4.4.19). This will have caused American judges to report the error considerably more frequently than other respondents. Admittedly, L-vocalisation is also quite common with other English speakers (see 4.2.19), but judging by the dramatically low detection rates in the British

Isles and the southern hemisphere, the latter phenomenon is not subject to the same stigmatisation as in the United States. The intended error in TELL was certainly not detected by any judges from Australia or New Zealand, possibly since these accents have also been described as having either a pharyngealised or vocalised [t̪] (see 4.2.19). While different authors mention the stigma attached to L-vocalisation in RP or other varieties of British English (e.g. Collins & Mees 2003b: 169, Wells 1982: 314), the present results show that this stigmatisation has either decreased or is at least not as applicable to foreign-accented speech as has previously been assumed. This means that Jenkins (2000: 138–139) may well be justified in assigning a low priority to “the production of dark [t̪]”, especially with regard to British English. The same conclusion is drawn by Wells (2005: 105) in his review of Jenkins’s recommendations. In fact, there is considerable evidence to indicate that the nature of post-vocalic /l/ is changing in Britain and that vocalic realisations are increasingly frequent among speakers of different social backgrounds (see 4.2.19; Tollfree 1999, Wells 1997, 2005).

3.5.21 Assessment of COLOUR

The overall composite severity estimate for COLOUR warrants a placing of this token in the upper-intermediate range of significant errors (3.2.2); indeed, the RP estimate was even among the six highest (3.2.4), while the GA estimate only ranked among the intermediate errors (3.2.5). In fact, the estimates in the two versions differed by more than one Likert scale point. This dramatic difference can largely be ascribed to a disparity in hit rates. While 94% of judges in the RP version reported the error, this was only 28% in the GA form. (The difference was statistically significant, as may be seen in Table 3.73 – which also shows a similar divergence for strict and lenient judges.) Those judges who had reported the error, however, did not evaluate it significantly more strictly in either version of the experiment (although the AS estimates for women, younger and stricter respondents were in fact higher), which suggest that detection, not assessment, is a key factor in accounting for the inter-version differences for this token.

The difference between versions is also apparent in other ways. A breakdown of the composite severity estimates, as in Table 3.74, reveals that all significant variation is between groups taking part in different forms of the experiment. In addition, the error prompted 15% of RP respondents to make relevant observations, as opposed to 10% in the GA form. In fact, some of the latter were more concerned with the use of the British spelling of *colour* in the GA version than with details of pronunciation.

While it may be noted that the composite severity estimates are significantly higher for men (3.3.2) and for younger respondents (3.3.3), the most striking pattern in the evaluation of this token is clearly the variation between different versions of the experiment. The high detection rates in the RP version are largely consonant with the importance attributed to COLOUR in Collins *et al.* (1987: 95) and Collins & Mees (2003b: 291), partly as a result of confusion over spelling (Collins *et al.* 1987: 59, Collins & Mees 2003b: 95; see also Gimson & Cruttenden 1994: 104). According to Brown (1988: 222), the contrast between

/ʌ/ and /ɒ/ actually has one of the highest functional loads of all pairs of vowels in RP. While the error is not explicitly described as significant in either Dretzke (1985) or Jenkins (2000), Gussenhoven & Broeders (1997: 99) point out that confusion between /ʌ/ and /ɒ/ is also “due to exposure to models other than RP, in particular American English accents, in which RP /ɒ/ frequently corresponds to the unrounded vowel /ɑ:/. This may cause learners to use Dutch /ɑ/ instead of /ʌ/. It is, however, unclear, to what extent this would prompt Dutch students of GA to use Dutch /ɑ/, GA /ɑ:/ or even RP /ɒ/ in those cases where GA has /ʌ/, as in *color*. Nevertheless, Collins & Mees (1993: 71, 128) explicitly warn students of American English about confusing /ɑ:/ and /ʌ/. If the present results are to be depended on, there appears to be considerable tolerance of variant realisations of /ʌ/ in American English. This is regardless of whether, in the word *color*, the GA actor is perceived as saying either /ɑ/ or /ɒ/.

Table 3.73. Hit rate and adjusted severity coefficients for COLOUR, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald Z}| \geq 2$ (in **bold**).

COLOUR	Sex	Age	Leniency	Version
Hit rate coefficient	0.001	-0.010	0.783	-3.320
Standard Error	0.233	0.009	0.074	0.265
Wald Z	0.004	-1.111	10.581	-12.528

COLOUR	Sex	Age	Leniency	Version
Adjusted severity	0.408	-0.014	0.910	0.047
Standard Error	0.113	0.005	0.025	0.155
Wald Z	3.606	-3.049	36.864	0.304

Table 3.74. Severity estimates for COLOUR, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	3.178	0.095
GB/NRP	3.137	0.104
IRL	3.210	0.207
AU&NZ&SA	2.521	0.229
US/GA	1.037	0.167
US/NGA	0.941	0.154
CDN	1.030	0.225

It is not as if all American judges failed to observe the loss of a phonemic contrast between *color* and *collar*. One respondent from the West Coast even observed that “this one is comedy – he’s collar blind instead of c[o]lour (OK, Brit spelling) blind” (Subject 696). It is, however, possible that Dutch substitutions such as /a/ or /ɒ/ may be within the range of tolerance for some speakers. Speakers whose own accents conflate /a/ or /ɒ/, for instance (Wells 1982: 476), may be more prone to this. A similar leniency may be associated with respondents involved in the Northern Cities Chain Shift who realise /ʌ/ as [ɔ], as was suggested in 4.2.20. Unfortunately, as was pointed out in 4.6, no conclusive evidence was found to show that those who use [ɔ] for /ʌ/ are more tolerant of similar substitutions in Dutch English. One can only conclude that, on the basis of the results for COLOUR, conflation of [ɔ] for /ʌ/ is much less likely to be detected by North Americans than by other listeners. It is, however, still an important priority for learners of RP, and possibly other varieties of British, Irish or Antipodean English. After all, only three out of 33 Irish participants did not detect the error, and only six out of 33 listeners from New Zealand, Australia and South Africa.

3.5.22 Assessment of STOOD

According to Collins & Mees (2003b: 290), the confusion of /ʊ ~ u:/ and the “articulation of /ʊ/” are among the “most significant” and “persistent” errors of Dutch learners of English (see also Collins *et al.* 1987: 95, Collins & Mees 1993: 128). Gussenhoven & Broeders also cite these as features of Dutch English (1997: 97) but do not proscribe them explicitly in their “Hints for the future teacher” (16). There is no mention of these errors in Dretzke (1985) or Jenkins (2000), but Brown (1988: 222) accords conflation of /ʊ ~ u:/ a low rank ordering in terms of functional load.

Apart from these varying assessments of the importance of the error in STOOD, it should be noted that some native speakers of varieties of British English, notably in Scotland and Northern Ireland, do not make a phonemic contrast between /ʊ/ and /u:/, whereas this is not true of any North Americans – at least not in this phonetic environment (see 4.2.21 and 4.4.21). This would suggest that the token was judged more leniently by some listeners in the RP version. This was, however, not the case at all.

While the composite severity estimate for this token ranked among the upper intermediate in the overall hierarchy of error (see 3.2.2), and among the intermediate in both the RP and GA versions (see 3.2.4 and 3.2.5), it was in fact significantly higher in the RP form. As a matter of interest, the difference was more than one Likert scale point (3.2.6). Higher severity estimates were also found for men (3.3.2) and younger respondents (3.3.3). In addition, as Table 3.75 shows, all variations between major accent groups taking part in different versions were significantly different.

Table 3.75. Severity estimates for STOOD, broken down by major accent group.

Major accent group	Estimate	Standard Error
GB/RP	2.861	0.126
GB/NRP	2.994	0.122
IRL	3.006	0.228
AU&NZ&SA	2.781	0.225
US/GA	1.197	0.170
US/NGA	0.974	0.152
CDN	0.748	0.220

A breakdown of the composite severity estimates into hit rates and adjusted severity reveals that inter-version differences must be attributed not to the importance attached to this error by different judges, but by their ability to detect it. As Table 3.76 shows, the AR estimate for North Americans is not significantly different, but their HR estimate is significantly lower. In addition, younger and self-identified stricter judges have significantly higher HR and AS estimates than older and more lenient judges, which is clearly in keeping with younger listeners' higher composite severity. (Such effects of age on detection and severity are discussed in 3.4.4.) While sex significantly affects the AS estimates (as has been observed above), this is not true of the HR estimates. If women judge the errors they have reported significantly more strictly than men, and if there is no statistically significant difference in detection between men and women, this would suggest that women should have a higher rather than a lower composite severity estimate. This is not the case. In fact, only 32% of male participants failed to detect the error, as opposed to 44% of female listeners. If these differences are not significant in a multi-level model, this may well be because a great deal of this variance is subsumed under other factors. In other words, the divergence in detection rates for men and women can be accounted for by factors other than sex – such as age, self-identified leniency and, most notably, version.

Table 3.76. Hit rate and adjusted severity coefficients for STOOD, broken down by sex, age, leniency and version. Significance is obtained if $| \text{Wald Z} | \geq 2$ (in **bold**).

STOOD	Sex	Age	Leniency	Version
Hit rate coefficient	-0.151	-0.025	0.594	-2.766
Standard Error	0.214	0.008	0.061	0.231
Wald Z	-0.706	-3.125	9.738	-11.974

STOOD	Sex	Age	Leniency	Version
Adjusted severity	0.420	-0.020	0.898	0.243
Standard Error	0.122	0.005	0.027	0.168
Wald Z	3.437	-2.340	33.514	1.450

There is no denying that for STOOD, inter-version variation in detection rates is particularly striking. After all, 88 of listeners in the RP form reported the error, as against a mere 26% of North Americans. This need not only be accounted for in terms of accent variation. Interestingly, 25% of GA listeners (as against one single listener in the RP form) selected the word “still” as an error, locating it mostly in the initial cluster /st/. Their comments make clear that they were mostly concerned with the absence of /t/, something which did not actually become immediately apparent from a post-hoc inspection of the auditory stimuli. Listeners stated variously that “the **t** in ‘still’ is almost entirely missing” (Subject 808) and that “[u]nlike the previous error (**th**), which can be attributed to the lack of a **th** sound in many languages, the missing **t** in “still” will be much harder to understand” (Subject 702). Controversially, some claimed either that /t/ should have been aspirated more (Subject 699) or that it should have been as “hard” as in “bat” (Subject 522). This attention to “still” may be ascribed to the GA actor’s performance, but it is also possible that STOOD’s lack of salience to some North American judges caused them to look elsewhere for errors. Whether or not these respondents were distracted by the realisation of the onset of “still”, the fact remains that STOOD was reported much less in the GA version, and not more, as may be expected given the conflation of /ʊ ~ u:/ in some Northern British and Irish accents. (In fact, as will be shown in 4.6, respondents who are likely to have such mergers in their own speech were demonstrably *less* lenient of the error in STOOD.) Nevertheless, it is crucial to note that, since the AS estimates were the same for both versions, those listeners who detected the error in the RP version did not judge it significantly differently than those who reported it in the GA form. In other words, the conflation of /ʊ ~ u:/, once identified, is equally important to both groups. All this would seem to argue against Wells’s suggestion (2005: 106) that, since “[m]illions of Scottish speakers of English manage perfectly well without any difference between the vowel of *shoot* and that of *foot*” and as this distinction has a “low functional load”, it is not required in English as an international language.

The equal significance attached to STOOD was also evident from the 66 observations prompted by the error in STOOD. The percentage of judges providing relevant comments was similar in both versions: 11% in the GA version and 12% in the RP form. 56 of these observations were couched in fairly neutral terms, explaining the source of the error in terms of either vowel quantity, quality or stress. Some pointed out that North Americans could interpret this realisation of *stood* as *stewed*, which would not be possible in other relevant accents of English. While there was only one dismissive American comment

(“Perhaps *stood* is too long, but I wouldn’t classify it as an error” – Subject 314), there was not a single negative one. Conversely, the RP version saw three negative comments (including the arguably condescending “It is endearingly Dutch” – Subject 374), but also seven that were dismissive – sometimes because respondents related it to the conflation of /ʊ ~ u:/ found in certain accents. This is illustrated in Table 3.77.

Table 3.77. Dismissive comments on STOOD in the RP version.

Subject	Accent self-identification	Comment
331	British standard, Northern accent	/u:/ in “stood”. Again, lots of accents do this, and it sou[n]ds like a mixed accent.
676	South African	I would have preferred to say “marked” rather than serious.
887	South African	Not sure if this counts as clearly detectable though
951	British (more or less RP)	/stu:d/ is not RP but I think OK
966	Northern British	Delightful, and not far from a pronunciation in Yorkshire dialect!
1005	North East Scottish (Aberdeen/ Dundee area) but softened in recent years	To me the “oo” in “stood” is not pronounced as in RP but I could not describe this as an error.

3.5.23 Assessment of INT1, INT2, INT3

As Dalton & Seidlhofer (1994: 75) point out, “intonation is a crucial element of verbal interaction, and most authors of teachers’ handbooks and teaching materials agree on this”. As they state, some of these writers actually stress the priority of suprasegmental phenomena over segmentals, even though others consider intonation to be unteachable. Jenkins, for instance, only includes limited aspects of intonation in her “Lingua Franca Core” (2000: 151). Nevertheless, many others, including Gimson & Cruttenden (1994: 276), have accorded intonation a high priority. For example, Dretzke (1985: 207), in his study of native English reactions to German-accented English, places intonation, rhythm and weak forms in the fourth highest “Dringlichkeitsstufe” when it comes to teaching English pronunciation to Germans. Similarly, Gussenhoven & Broeders (1997: 17) point out that, for Dutch learners of English, the importance of the attitudinal aspect of intonation “is very considerable indeed”. Collins & Mees (2003b: 291) classify “[m]onotonous intonation owing to restricted intonation range and lack of high heads” as found in Dutch English as a “significant error”. In fact, intonation is among the areas most frequently discussed in the pronunciation materials surveyed by Wrembel (2005: 428).

The importance ascribed to intonation in textbooks is less evident from experiments in which native speakers are asked to assess the seriousness of

specific intonational errors as against segmental or other types of suprasegmental errors. Dretzke (1985: 175), for instance, found that German “Sägeblattintonation” was not considered to be particularly serious by his Northern English respondents; similarly, Koster & Koet (1993: 76) noted that both Dutch and English respondents reported only few intonation errors when judging tape recordings of Dutch English sentences.⁷ This is possibly connected to what Vaissière (2005: 253) describes as

the difficulties that researchers face when they approach the study of intonation: lack of clear definitions, non-applicability of otherwise standardized experimental methods used in psychoacoustics and laboratory phonology, the effects of phonetic and melodic contexts and the speakers’ native language on the perception of intonational phenomena.

Such complications may be responsible for a divergence between native speakers’ assessments of intonational errors and the significance frequently attached to intonation in textbooks aimed at foreign learners. In the present experiment, the severity scores accorded by respondents to INT1, INT2 and INT3 were among the lowest of all tokens. While this could suggest that intonational errors (or at least those selected here) are less important than is often assumed, these results could also be due to the difficulties described by Vaissière. However, while Vaissière may be justified in referring to the “lack of standardized methods” in intonation research (Vaissière 2005: 241), it should be pointed out that the superposition technique used in resynthesising carrier sentences can be considered a tried and tested method. For instance, in a number of experiments on Dutch intonation in English, Willems (1982: 148) found this technique to be “an excellent manner for eliciting consistent acceptability judgements from native English speakers”. Although this fails to address Vaissière’s other concerns, it does at least suggest that a closer inspection of the three carrier sentences, and their different intonation patterns, need not be a purposeless exercise.

As is apparent from both the overall hierarchy of error (see 3.2.2), respondents considered the intended error in INT1 to be significantly less severe than those in INT2 and INT3. This was true for both the RP and GA versions. In addition, tokens INT2 and INT3 did not differ significantly from each other in either version of the experiment (see 3.2.4 and 3.2.5). In the GA version as well as in the overall hierarchy of error, INT1 was rated the least important of all 32 errors, and only TELL came below it in rank in the RP form. None of these tokens discriminated for sex (see 3.3.2.), but INT2 was assessed significantly more severely by younger respondents (see 3.3.3.) and North Americans (see 3.2.6). The differences by major accent group were not significant (see Table 3.78). In sum, INT1 was considered the least serious of all intended

⁷ According to Schuderer (2002: 16), the German use of “sawtooth intonation” (Sägeblattintonation) in English is caused by unaccented syllables descending too sharply between stressed syllables or at the end of a phrase.

intonation errors, with INT2 and INT3 being assessed as marginally more significant. In addition, INT2 was the only one to be evaluated significantly more strictly by any particular group of respondents.

Table 3.78. Composite severity estimates for INT1, INT2 and INT3, broken down by major accent group.

Major accent group	INT1		INT2		INT3	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
GB/RP	0.161	0.056	0.550	0.104	0.720	0.121
GB/NRP	0.156	0.065	0.305	0.090	0.643	0.124
IRL	-0.015	0.039	0.471	0.196	0.669	0.233
AU&NZ&SA	0.208	0.121	0.348	0.155	0.505	0.198
US/GA	0.062	0.043	0.691	0.133	0.863	0.156
US/NGA	0.159	0.072	0.651	0.126	0.883	0.142
CDN	0.060	0.077	0.513	0.185	0.539	0.189

While INT1 was only reported by 6% of respondents (6% in the RP version and 5% in the GA version), INT2 was reported by 19% (17% for RP and 23% for GA) and INT3 by no fewer than 27% (25% for RP and 32% for GA). A breakdown of the composite severity estimates into hit rates and adjusted severity coefficients is provided in Table 3.79. In all cases, a significant part of variation may be ascribed to self-identified leniency, and in the case of AS estimates to the almost predictably higher severity scores for women. This, however, does not apply to variation by version, not even in the case of INT2.

It is perhaps not very surprising that the inter-version difference in assessing INT2, as found for the overall severity estimate, should no longer be significant when broken down into HR and AS estimates. After all, the estimated composite severity for this token is only marginally higher (0.24 Likert scale unit) for North Americans than for listeners in the RP version, and the effect, though significant, is in fact quite small ($\chi^2 = 6.26$, $df = 1$, $p < .05$; see 3.2.6). In addition, some of the variation for INT2 must be ascribed to younger respondents' significantly higher HR estimates.

Not only is it difficult to account for the varying assessment of INT2, but the differences in ranking between the three intonation tokens are also hard to explain. For example, it may be somewhat puzzling that so few respondents

reported or detected INT1 – as compared with INT2 and INT3. Dutch pitch contours had been superimposed on all three carrier sentences, yet clearly the effect of this was not equally salient in all cases. While an overwhelming majority of judges did not appear to have any problems interpreting the three different intonation contours as falling within the range of acceptable variation, a minority reported errors both for INT3 and INT2 – though only rarely for INT1.

Table 3.79. Hit rate and adjusted severity coefficients for INT1, INT2 and INT3, broken down by sex, age, leniency and version. Significance is obtained if $|\text{Wald } Z| \geq 2$ (in **bold**).

INT1	Sex	Age	Leniency	Version
Hit rate coefficient	-0.334	-0.001	-0.911	0.486
Standard Error	0.360	0.014	0.100	0.393
Wald Z	-0.928	-0.071	-9.110	1.237

INT1	Sex	Age	Leniency	Version
Adjusted severity	0.821	0.008	0.7981	-0.141
Standard Error	0.316	0.015	0.085	0.371
Wald Z	2.596	0.520	9.388	-0.379

INT2	Sex	Age	Leniency	Version
Hit rate coefficient	-0.369	-0.031	-0.452	0.266
Standard Error	0.211	0.009	0.054	0.216
Wald Z	-1.749	-3.444	-8.370	1.232

INT2	Sex	Age	Leniency	Version
Adjusted severity	0.500	-0.015	0.688	0.299
Standard Error	0.170	0.008	0.045	0.170
Wald Z	2.941	-1.940	15.339	1.758

INT3	Sex	Age	Leniency	Version
Hit rate coefficient	-0.130	-0.010	-0.320	0.221
Standard Error	0.186	0.007	0.047	0.192
Wald Z	-0.699	-1.429	-6.809	1.151

INT3	Sex	Age	Leniency	Version
Adjusted severity	0.375	-0.005	0.736	0.259
Standard Error	0.173	0.007	0.042	0.168
Wald Z	2.167	-0.783	17.618	1.545

Post-hoc inspection of the auditory stimuli by means of the contour analysis often used in British intonation research (Cruttenden 1997: 38) confirms that all three sentences may easily be perceived as having plausible intonation patterns in native English. As Cruttenden points out, “it is generally a truism that **almost** any tone can be used in any context” (Cruttenden 1997: 118, author’s emphasis). However, it should be added (as Cruttenden also appears to suggest) that some of these patterns may be more likely to occur than others. For instance, the Dutch intonation used in INT3 may be interpreted as (1), with a mid-level head and a fall-rise on the nucleus, as opposed to traditional RP (2), with a high head and a low rise.⁸

(1) > *Are you taking the ~ car ?* ||

(2) *Are you ¹ taking the ,car ?* ||

According to Cruttenden, the latter pattern “seems in many ways to be the most neutral tone” in yes/no questions (1997: 104). Whereas a mid-level head would not be at all unusual in modern RP or similar varieties, and GA may favour a low rise on *are*, the introduction of a fall-rise, however, may be taken to imply “reservation and doubt” (Collins & Mees 2003b: 265). This change of directionality from low rise to fall-rise will sound more marked to some native-speaker judges, especially if they are not provided with a suitable context. This was quite evident from the 36 relevant comments, some of which actually described a context in which the pattern would be more fitting: “The meaning is confused. It sounds as if the question means *Are you taking the car or something else?*” (Subject 321). A South African participant suggested that the pattern “might be appropriate if the speaker is hostile” (Subject 887). In addition, at least 11 listeners mentioned that they had expected a rise or a more obvious rise. As Subject 584 stated, “Questions usually rise in tone at the end”. At the same time, the relatively low detection scores must serve as a reminder that for a great many respondents, any such deviation had gone unnoticed.

Cruttenden (1997: 109–110) points out that “some dialects of English (for example, the North-West Midlands accent of Staffordshire, West Derbyshire, Cheshire, and South Manchester) use fall-rises on interrogatives very frequently, while R.P. uses them relatively infrequently”. Participants from those areas, as well as anyone used to these accents, may not have been particularly struck by

⁸ The contour-type system of intonation analysis familiar from the British tradition (e.g. O’Connor and Arnold 1973; see also Cruttenden 1997: 38) is used here.

the use of the fall-rise in a yes/no question – unless stigmatisation is involved. To some North American judges, however, the intonation used in INT3 did sound quite marked: no fewer than five US respondents stated that it reminded them of British English (see Table 3.80). Not only did one American refer to a “European lilt” (Subject 198), but another stated that the intonation was “completely foreign-sounding” (Subject 812). Needless to say, this is not statistically significant, nor was this kind of reaction attested for the other two intonation tokens. Nevertheless, a point for future investigation would be to establish to what extent, if any, Dutch intonation patterns in American English are more generally perceived as “British”. It would also be interesting to discover if this is rated as a positive or a negative feature.

Table 3.80. North American reports of British intonation patterns in INT3.

Subject	Accent self-identification	Comment
92	American/Midwest	Sounds like British intonation
125	American/Standard	This sounds like Bri[t]ish intonation – American intonation on "car" would be different
285	American/Standard	Intonation should rise at the end; the sample sounds British
547	American/East Coast	This sounds like a British English intonation
696	US West Coast	This time the speaker's intonation veered off toward Britain.

Perhaps INT3 was reported more frequently than the other intonation errors because of the change of directionality (i.e. a fall-rise instead of a low rise). This explanation would be in keeping with the significance attached to this type of error in Willems's research into native-speaker reactions to the use of various Dutch intonation patterns in English. Willems (1982: 125) refers to Dutch “deviations” in directionality as “very relevant for the perception of non-nativeness”. It must be admitted, however, that in Willems's experiment, this relevance is based on a different change in direction made by Dutch learners. In his work, it is falls, not rises, that are being replaced by “a simple rise or a more complex movement ending in a rise” (Willems 1982: 110). If Willems's strictures can be generalised to include other changes in the direction of nuclear tones, this could help to explain the relative importance attached by both GA and RP respondents to INT3.

In Willems's hierarchy, a “deviation” possibly exemplified by INT2 is described as “continuation”, ranking below “direction” as being no more than “mostly relevant” (Willems 1982: 125). He describes “continuation” as the replacement of “complex rise-fall-rise movements, ... often used ... to mark

a prosodic boundary, by simple falls or rises” (Willems 1982: 110). This would appear to apply to the substitution of a fall-rise, as may be expected in traditional RP (3), by a simple rise, as in (4).

(3) *I \ didn't ° actually ° think that was ˇ true | but you ' may be ` right. ||*

(4) *I \ didn't ° actually ° think that was ´ true | but you ' may be ` right. ||*

The latter is a likely interpretation of the Dutch pitch contours used in INT2. As Collins & Mees (2003b: 264, 282) point out, the “over-use of high rise” may be considered “a Dutch error”, but in this case it could easily be used by a native speaker of English to “add marked emphasis”. As Subject 951 comments, “The sentence intonation seems a little unlikely, but in context, who knows?” This may explain why relatively few respondents reported INT2 as an error. It would also give additional support to Willems’s claim that deviations in directionality rank above those in continuation (Willems 1982: 125–126). However, it should be noted that, of the 21 relevant observations, two explicitly referred to a noticeable rise on *true*. There were also three technical comments to the effect that the carrier sentence sounded “computer-generated” (Subject 496) – a claim that was made only once for INT3 (“Sounds computer-generated, but very good” – Subject 496) and not at all for INT1. Such observations would help to indicate why, even if this could easily be interpreted as a plausible English pattern, it was nonetheless more salient than INT1.

The pattern used in INT1 could be interpreted as (5), with a high rise in the first intonation group, followed by a high head and a high fall in the second. Although it would be atypical of English to have a marked pitch change on a function word such as *they*, and this would tend to allot it undue prominence, it would be perfectly normal in English if the speaker actually wished to give additional emphasis to this word. As one respondent from the American West Coast suggested, “*They think* would normally receive less stress” (Subject 404). A more neutral pattern in, for instance, traditional RP would be (6), as a single intonation group with a low pre-head for *they think*.

(5) *´ They ° think | it's ' totally ` stupid. ||*

(6) *They ° think it's ' totally ` stupid. ||*

In addition, the more extended pitch pattern found in INT1 on *they*, causing it to start lower and end higher than would be expected in a more neutral version, may be similar to what is identified by Willems as “outset”. He describes this intonational “deviation” in Dutch learners as a tendency “to start at the Low level”, while native speakers of English generally start “an utterance on the Mid level” (Willems 1982: 111). Since so few listeners reported this token, and only nine commented on it, this would suggest that, at least in this case, “outset” is

not even “occasionally relevant” (to quote Willems), but hardly relevant at all (Willems 1982: 126).

It is interesting that Willems’s hierarchy for “direction”, “continuation” and “outset” appears to be reflected in the frequency with which INT3, INT2 and INT1 were reported and in the strictness – at least to some extent – with which they were assessed. Nevertheless, the fact remains that none of the intended errors were considered at all severe by respondents in either version of the experiment. Arguably, this could be taken to mean that the intonation of Dutch learners is less significant than certain segmental errors. However, it could just as easily be seen as a result of the selection of these particular patterns, or the difficulties inherent in testing reactions to Dutch intonation patterns superposed on English carrier sentences, especially if no appropriate context has been provided. In addition, the effect of repeated deviation, and its combination with other errors, may be particularly strong in the case of intonation. (The example of the moving walkways at Amsterdam Airport incessantly and unidiomatically exhorting travellers to *mind your step* springs to mind.) However, the design of this experiment did not allow for such effects to be tested. Clearly, more research of this nature will have to be done before intonation can be declared a low priority in pronunciation teaching to Dutch learners of English.

3.6 Comparison with severity assessment in the Dutch Experiment

As was pointed out in 1.2.3, previous research indicates that native judges tend to evaluate foreign learners’ errors considerably more leniently than non-native speakers, and that the former also tend to prioritise different types of errors, for instance “global” errors that “affect overall sentence organization” (Ellis 1994: 66, see also Dulay et al. 1982: 191). In the context of pronunciation, several comparisons of native and non-native judges have shown the former to attach more importance to prosodic rather than segmental errors (see Johansson 1978: 9–15, 123). However, Koster & Koet (1993: 89) have found that, while the native judges tended to “find fault with fewer vowels and consonants” than their Dutch respondents, both groups considered prosody to be relatively unimportant with the exception of word stress (a result that was also attested in the present Native-speaker Experiment.) While Johansson recommends that “non-native teachers of English should change their priorities in the area of pronunciation and attach more importance to prosody” (1978: 123), Koster & Koet (1993: 89–90) conclude that “current teaching practice in Holland is correct in paying hardly any attention to suprasegmental features, but that it is wrong in paying attention mainly to phonemic aspects” – the latter suggestion also being found in Johansson (1975: 82).

Such assertions may be tested by means of a comparison of the present Native-speaker Experiment with the earlier experiment involving Dutch university students, lecturers and secondary school teachers. (For a description of the experiment conducted in the Netherlands, and the groups which participated in this, see 2.1.) It would be of particularly interest to see if the teachers among these are stricter judges of Dutch pronunciation because of “the whole learning and teaching context against which he or she will inevitably view the errors” (Davies 1983: 310; see also Hughes & Lascaratou 1982). For instance, Koster & Koet (1993: 69) ascribe Dutch teachers’ greater objections to a Dutch accent to “undue fastidiousness”. Interestingly, Bongaerts (1999a: 9) found that native speakers were more reliable judges of pronunciation than non-native speakers when it came to identifying speakers as non-native, regardless of whether the former had experience of judging or teaching pronunciation.

The problem is, however, that the two experiments are not strictly comparable. Partly as a result of this, no direct evidence has been found to support the above claims. What the comparison does show, however, is that the evaluation of different types of errors by disparate groups of judges is subject to certain complicating factors, including a tendency on the part of non-native judges to underestimate the importance of a number of both phonemic and sub-phonemic errors, and the differences between RP and GA in this respect.

As became clear in Chapter 2, the earlier experiment with Dutch respondents differed considerably from that involving native-speaker judges. Most importantly, instead of being asked to both identify and assess pronunciation errors in audio recordings of complete sentences (with only one distractor), Dutch respondents were invited only to evaluate a number of real and imaginary errors on the basis of verbal descriptions using one-word examples (see 2.4.1). In addition, no explicit reference was made to the importance of particular errors with regard to different pronunciation models such as RP or GA (see 2.1.3). Moreover, only those errors considered to be the most significant in a specific category such as “phonemic” or “suprasegmental” were included in the Native-speaker Experiment (with a few exceptions and additions; see 2.4.3). Furthermore, the Dutch respondents were not asked to describe their own accents, estimate their own leniency or indicate their precise age (other than by age group); they were invited to take part in different versions of the experiment on the basis of their professional / educational background (e.g. students as opposed to teachers or lecturers) rather than on the variety of English they felt most competent to judge (RP or GA). These dissimilarities between the two experiments should be viewed within the context of their different aims – the Dutch Experiment being used, amongst other things, to pre-select tokens for the subsequent Native-speaker Experiment. Some of these differences have also arisen as a result of practical considerations and accumulating insight on the part of the researcher.

In brief, the different design of the experiments renders it difficult to compare the results reliably. To the extent that the results can be seen to be at all compatible, it should first of all be noted that it is only the adjusted severity

estimates of the native-speaker listeners that may be compared with the Dutch judges. This is because the composite severity estimate as used in the Native-speaker Experiment is partly composed of the error detection success rate (see 3.4.4). There is no direct equivalent in the Dutch test, where the intended errors had been mixed with distractors (see 2.1). It will be recalled that the difference between the composite severity estimate for a particular token and the adjusted severity can be quite dramatic. For instance, the adjusted severity estimate for TELL is 2.18 in the RP version, but its corresponding composite estimate is close to zero. This difference makes it harder to discuss the Dutch participants' judgements in the context of the hierarchies of error as presented in 3.2.2, 3.2.4 and 3.2.5, since these are based on composite severity estimates. Secondly, it is only those 22 tokens that are very similar or virtually identical that can serve as a basis for comparison (see 2.4.3), which excludes four tokens illustrating TH-stopping in medial and final position (WEATHER, BREATHE, AUTHOR and BOTH), none of which featured in the Dutch Experiment. Neither does it include five tokens illustrating specific suprasegmental issues (THAT_THA, WOULD_ON, INT1, INT2, INT3), which had not been provided with actual examples in the Dutch Experiment. This makes it difficult to verify the claim that non-natives may underrate the effects of prosody (as compared with native speakers). The distractor INDIA, which is dissimilar in the two experiments, has also been excluded. Thirdly, factors such as age, leniency and major accent group should not be included in any estimation of the various different coefficients.

The MLwiN program was used to calculate the severity estimates of the relevant 22 tokens according to "version of the experiment", both individually and overall. In the case of the Native-speaker Experiment, "version" has been used to refer to the RP and GA forms respectively, whereas in the Dutch Experiment this refers to the professional/educational background of the respondents. To take just one example, comparing these different versions of the two experiments may give some indication of whether the error assessments of teachers of English in the Netherlands are similar to those of judges in North America.

The overall severity estimates for 22 selected tokens are presented in Table 3.81. Pairwise comparisons among versions revealed that, after Bonferroni adjustment for multiple comparisons among $k = 5$ group means, only three of the differences between these versions did *not* reach significance: GA and NL/STU ($\chi^2 = 3.23$, $df = 1$), RP and NL/SST ($\chi^2 = 3.29$, $df = 1$) and GA and NL/SST ($\chi^2 = 0$, $df = 1$). In so far as it is indeed possible to compare both experiments, this suggests that, where these selected tokens are concerned, the judgements of Dutch secondary school teachers did not differ significantly from the adjusted severity estimates of either group of native speakers. While Dutch students' assessments appear to be similar to the adjusted severity (AS) estimates of North Americans, this does not appear to be true of those of listeners in the RP version, nor of any comparisons involving the Dutch lecturers.

Table 3.81. Overall severity estimates, broken down by version in the two experiments.

Versions of the two experiments	Estimate	Standard Error
RP	3.284	0.059
GA	3.489	0.065
NL/LEC	3.742	0.088
NL/STU	3.746	0.082
NL/SST	3.384	0.074

One could possibly infer from this that Dutch lecturers in English are less in tune with native-speaker pronunciation judgements than Dutch secondary school teachers of English, and that Dutch students of English agree more with North Americans than those in the RP version of the experiment when it comes to assessing the severity of selected pronunciation errors. The latter in particular may be considered a reflection of what Van der Haagen (1998: 102) has found to be “a quantifiable American component in the English pronunciation of Dutch secondary school pupils”. However, it would also imply that, since North American AS estimates tend to be a little higher than those of RP listeners, a corresponding strictness is to be found in Dutch students – even though these students themselves believe judges from the British Isles to be stricter than Americans and Canadians (see 1.1). Be that as it may, it should be borne in mind that the restrictions that any comparison of these two experiments are subject to continue to apply here as well – particularly as regards the distinction between adjusted and composite severity. Moreover, a different picture emerges when the individual estimates for all 22 tokens are calculated (see Table 3.82).

Table 3.82. Severity estimates of 22 selected tokens in the two experiments, by version.

	RP	Standard Error	GA	Standard Error	NL/LEC	Standard Error	NL/STU	Standard Error	NL/SST	Standard Error
BED	3.35	0.096	3.72	0.107	4.75	0.113	4.67	0.136	4.26	0.121
BAT	3.48	0.104	3.72	0.116	4.72	0.122	4.61	0.142	3.97	0.129
VAN	3.42	0.106	3.71	0.118	4.53	0.133	4.47	0.148	4.00	0.133
WINE	3.23	0.108	3.36	0.119	4.49	0.134	4.12	0.168	3.87	0.138
THIN	3.70	0.104	3.62	0.116	4.38	0.145	4.43	0.147	3.90	0.138
OFF	2.87	0.129	2.92	0.135	4.32	0.161	3.41	0.200	3.13	0.180
THAT	2.67	0.135	2.84	0.136	3.84	0.169	4.03	0.167	3.35	0.155

RED	3.38	0.114	3.95	0.124	4.09	0.186	3.51	0.184	3.75	0.156
ICE	2.88	0.127	3.18	0.146	4.10	0.164	4.10	0.175	3.50	0.155
TIE	3.23	0.124	3.38	0.166	3.61	0.188	3.85	0.184	2.91	0.162
DEAD	3.60	0.113	3.92	0.120	3.56	0.183	4.22	0.170	3.20	0.166
FILM	3.42	0.122	3.60	0.129	3.99	0.193	4.36	0.179	4.13	0.159
CAR	2.62	0.113	2.72	0.127	2.31	0.190	2.91	0.196	2.74	0.164
HOT TEA	3.28	0.154	3.59	0.171	2.35	0.227	2.10	0.209	2.30	0.194
NEW	2.80	0.141	2.09	0.170	2.31	0.215	2.87	0.212	2.52	0.189
IMAGIN	3.67	0.095	3.88	0.105	4.45	0.130	4.27	0.142	4.33	0.200
PERFECT	3.82	0.095	3.87	0.104	4.46	0.135	4.33	0.149	4.29	0.123
TO WALES	3.43	0.102	3.75	0.111	3.47	0.202	3.66	0.158	3.60	0.149
SECONDAR	2.66	0.109	2.67	0.120	2.03	0.177	2.66	0.173	2.36	0.155
TELL	2.18	0.226	2.86	0.172	2.63	0.201	2.82	0.219	2.21	0.188
COLOUR	3.32	0.116	3.05	0.147	3.96	0.192	3.31	0.203	3.12	0.165
STOOD	3.31	0.121	3.19	0.143	4.03	0.176	4.09	0.173	3.47	0.159

It is already immediately apparent from Table 3.82 that, while no native-speaker estimates exceed 3.82 (in the RP version) or 3.95 (in the GA version), the highest Dutch estimates peak at 4.75 (Dutch lecturers), 4.67 (Dutch students) and 4.33 (Dutch secondary school teachers). Since the lowest estimates are 2.62 (RP version), 2.09 (GA version), 2.31 (Dutch lecturers), 2.10 (Dutch students) and 2.30 (Dutch secondary school teachers), this would suggest that the Dutch judges used a wider-ranging scale. This appears to be particularly true of the lecturers and the students. If the Dutch and native judges made different use of the Likert scale, this is quite likely to be the result of the relative dissimilarity between the two experiments. It also makes it difficult to interpret similarities and differences between particular groups' judgements of individual tokens with any degree of confidence. More specifically, it does not provide any support for categorical claims that non-natives are more severe than natives, or that those who teach English are stricter than those who do not.

The varying assessments of BED in different versions may serve as an example. Pairwise comparisons among versions revealed that, after Bonferroni adjustment for multiple comparisons among $k = 5$ group means, only one difference did *not* reach significance: GA and NL/SST ($\chi^2 = 0.39$, $df = 1$). This means that BED was assessed significantly differently in all pairwise comparisons of versions except this. If judges in the GA and NL/SST versions used the Likert scale differently, however, this would mean that any evaluations of individual tokens that happen to be equally high are more likely to point to differences than to similarities. While the assessment of BED is not significantly different in the NL/SST and GA versions in *absolute* terms, this token was indeed judged differently in *relative* terms. The latter can be illustrated by means of regression analysis of the various severity estimates.

Regression analysis can be used to predict the severity score of a particular group of Dutch judges from the adjusted severity score of a group of native speakers, and to identify those items that are well away from the resulting regression line, i.e. tokens that are judged differently by the two groups of respondents. To identify these items, an absolute residual of 1 Likert scale point was used as a criterion value. This has been done for all combinations of versions, using the 22 relevant severity estimates obtained through MLwiN. For each comparison, one table will give the regression coefficients (A and B) and the correlation coefficient (R), and a second table will provide the residual scores for these outlier tokens. Table 3.83, for instance, shows the regression coefficients for NL/SST as may be predicted from the GA version. The fact that the correlation coefficient of .694 is relatively strong makes it possible to identify the outliers reliably.

Table 3.83. Regression coefficients for NL/SST predicted from the GA version (R = .694).

	Coefficient	Standard Error
A	0.266	0.736
B	0.938	0.218

As can be seen in Table 3.84, BED has a residual higher than 1 scale point (1.019), which means that this item was assessed a little less leniently by the Dutch secondary school teachers than is to be expected from the general pattern established by respondents in the GA form. Conversely, Dutch secondary schoolteachers were somewhat less strict on TIE, DEAD and TELL (all with residuals lower than -1), and much less severe on HOT_TEA (which has a very low residual of -2.675), than would be expected from the correlation between the two groups. This suggests that BED *was* in fact judged considerably differently by the two groups – in addition to four other tokens.

Table 3.84. Outliers in the regression analysis of NL/SST and the GA version, using a criterion value of 1 Likert scale point for the standardised residual.

BED	1.019
TIE	-1.043
DEAD	-1.480
HOT_TEA	-2.675
TELL	-1.497

While a pairwise comparison of the GA and NL/SST severity estimates for all outliers provided in Table 3.84 actually shows (after Bonferroni adjustment for multiple comparisons among $k = 5$ group means) that the only token to be assessed significantly differently ($\chi^2 = 11.62$, $df = 1$) is not BED but DEAD,

regression analysis strongly suggests that it is actually all five that may be identified as having been judged differently from what the correlation between the two groups would lead one to expect – especially *HOT_TEA*. In other words, the notion that *BED* was not judged significantly differently in the NL/SST and GA versions does not tally at all with the results as provided by regression analysis.

Clearly, regression analysis provides a useful tool for detecting those tokens which were assessed lower or higher than may be expected on the basis of correlations between groups. It is of interest to note that in the various different comparisons between versions, some of the same tokens continue to emerge as outliers – even though the only token to be identified as such in all comparisons is *HOT_TEA*. This is also evident from a comparison of NL/SST and the RP version. Table 3.85 shows there to be an even stronger correlation coefficient of .730, which makes it possible to identify the outliers reliably in Table 3.86 as *BED*, *THAT*, *TIE*, *DEAD* and *HOT_TEA*. While the first and the last three of these are at a similar distance from the regression line as in the comparison between NL/SST and the GA version, this not true of *THAT*, which was judged a little more strictly by the Dutch secondary school teachers than would be expected from the correlation between these and judges in the RP form. Conversely, *TELL*, which is not identified as an outlier in this comparison, is judged less strictly by Dutch secondary school teachers than the correlation of their results with North American respondents would lead one to expect.

Table 3.85. Regression coefficients for NL/SST predicted from the RP version ($R = .730$).

	Coefficient	Standard Error
A	-0.385	0.799
B	1.186	0.248

Table 3.86. Outliers in the regression analysis of NL/SST and the RP version, using a criterion value of 1 Likert scale point for the standardised residual.

<i>BED</i>	1.435
<i>THAT</i>	1.191
<i>TIE</i>	-1.132
<i>DEAD</i>	-1.443
<i>HOT_TEA</i>	-2.544

Interestingly, Dutch teachers' and native speakers' different assessments of *THAT* and *TELL* appear to reflect variations in assessment between respondents in the RP and GA versions; not only are the AS estimates for these tokens a little higher in the GA version, but their composite severity estimates show that North Americans judged these tokens significantly more severely (see 3.5.6 and 3.5.20). To the extent that these results can be compared at all, this

would suggest that Dutch secondary school teachers share North Americans' somewhat stricter assessment of THAT (see 3.5.6), and the RP judges' more lenient evaluation of TELL (see 3.5.20).

Widdowson was perhaps the first to make the notorious claim that native speakers are irrelevant to the development of English in the world (Jenkins 2000: 7). Needless to say, this view is not generally accepted, either by native or non-native speakers of English (see, for instance, Scheuer 2005, Trudgill 2005b, Wells 2005; and 3.7). If, therefore, some account is taken of native-speaker judgements of Dutch pronunciation errors, the following could be concluded on the basis of regression analysis of the results of the two experiments. Purely based on AS estimates rather than composite severity, native speakers attach somewhat more importance to the lack of aspiration in TIE and the use of a glottal stop in DEAD than do Dutch secondary school teachers, and much more to degemination in HOT_TEA, but less to final devoicing in BED. (It is not entirely unlikely that these results may have been influenced by the fact that the Dutch Experiment, but not the native-speaker one, uses phonetic terminology such as "aspiration" and "glottal stop", which may have confused some respondents.) In particular, North American respondents consider a dark [ɮ] in TELL to be more significant, while RP judges tend to judge TH-stopping in THAT less severely. This is useful information in so far as it gives an indication as to which pronunciation errors, once detected, are prioritised more by particular groups of native speakers. It does not factor in to what extent some of these errors are more likely to remain undetected, even though this is clearly a very salient indication of an error's severity. If the composite severity estimates are taken into consideration, however, it turns out that native speakers attach considerable significance only to some of these tokens.

In the overall hierarchy of error (see 3.2.2), it is only BED and DEAD that are found in the upper-intermediate range of important errors (3.5–2.0). While this is the same for the upper range (> 2.0) in the GA hierarchy (see 3.2.6), in the RP hierarchy (see 3.2.5), TIE is the only other token of those mentioned above that has an estimate higher than 2.0. Relatively high composite severity estimates indicate that not only were these tokens assessed fairly severely, but that they could also be easily detected. Of these, it is only DEAD and TIE that the Dutch secondary school teachers had a slight tendency to underestimate. Since BED is among the most important errors detected and assessed by native speakers, one can hardly fault the secondary school teachers for overrating its significance. This suggests that Dutch secondary school teachers would do well to attach more importance to the correct use of glottal stops and, especially if their model is RP, to aspiration. They may also wish to consider to what extent their assessment of dark [ɮ], TH-stopping and degemination is consistent with the attitudes of particular groups of native speakers. As far as the other 16 tokens are concerned, teachers' priorities do not appear to differ strikingly from those of native-speaker judges – to the extent that the two experiments can of course be compared reliably, and detection rates are ignored.

It is not unthinkable that some Dutch secondary school teachers prioritised errors that unmistakably result in phoneme conflation and perhaps attached a little less importance to errors that appear to be realisational, such as TIE and DEAD. All the same, the failure to aspirate, or the glottal replacement of a lenis consonant, may well give the impression of phoneme conflation to native speakers – aspiration being particularly important if RP is adopted as a model (see 3.5.10), and the avoidance of glottal substitution in lenis stops even more so in the case of GA (see 3.5.11). Teachers of pronunciation should be aware that approaches that do not go beyond the phoneme may in fact do learners a disservice (as was also suggested in Johansson 1975 and Koster & Koet 1993). After all, phonetic features such as aspiration and the effect of vowel length on the fortis/lenis distinction are even included in Jenkins's (2001: 140) *Lingua Franca Core*, and their importance is frequently emphasised in pronunciation manuals intended for Dutch learners (see, for instance, Collins & Mees 2003b: 290–291, Gussenhoven & Broeders 1997: 16). Anyone engaged in teaching pronunciation to Dutch learners would probably do well to include such phonetic features in their training programmes, especially since failure to use these correctly is likely to result in phoneme conflation.

A slight tendency to underestimate the significance of the glottal replacement of lenis consonants was also found with the Dutch lecturers (but *not* the Dutch students) as compared to the AS estimates of judges in the RP and GA versions. This is evident from the regression analyses presented in Tables 3.87 to 3.90. In addition, the lecturers also appeared to attach too little importance to degemination (as did the other Dutch groups). They also considered quadrisyllabic realisations of words like *secondary* less significant than the correlation with the native-speaker judges' AS estimates would lead one to expect. The lecturers also had a slight tendency to underestimate the significance of NEW (as against the RP judges) and CAR (as against the GA judges). In view of this, it is tempting to assume that they may have underrated the effects of dialect mixing as illustrated by all three tokens.

It should be noted that the Dutch Experiment did not state which model of English the various pronunciation errors should be judged against (see 2.1.3), whereas this choice of model was made quite explicit in the Native-speaker Experiment. Since a four-syllable realisation of *secondary* is the standard pronunciation in GA, the lecturers may have considered it unjustified to accord it undue significance. Such a relatively low estimate for SECONDAR is, however, not evident from the results of the students and secondary school teachers.

In addition, the Dutch lecturers appeared to prioritise WINE and OFF a little more than the AS estimates in either native-speaker version of the experiment would suggest. While they also attached somewhat more importance to BED, THAT and ICE than would be expected on the basis of the relevant RP estimates, their assessments of these three tokens appeared to be consonant with those in the GA form (although they prioritised TO_WALES a little less than the latter group). In this context, it may be noted that especially the RP judges' *composite* severity estimates of THAT and ICE are also surprisingly mild, particularly

considering the strictures of pronunciation handbooks such as Collins & Mees (2003b).

Table 3.87. Regression coefficients for NL/LEC predicted from the RP version (R = .627).

	Coefficient	Standard Error
A	-0.407	1.163
B	1.299	0.361

Table 3.88. Outliers in the regression analysis of NL/LEC and the RP version, using a criterion value of 1 Likert scale point for the standardised residual.

BED	1.168
WINE	1.022
OFF	1.447
THAT	1.123
ICE	1.109
DEAD	-1.042
HOT_TEA	-2.176
NEW	-1.323
SECONDAR	-1.475

Table 3.89. Regression coefficients for NL/LEC predicted from the GA version (R = .617)

	Coefficient	Standard Error
A	0.184	1.027
B	1.065	0.304

Table 3.90. Outliers in the regression analysis of NL/LEC and the GA version, using a criterion value of 1 Likert scale point for the standardised residual.

WINE	1.044
OFF	1.477
DEAD	-1.149
CAR	-1.098
HOT_TEA	-2.373
TO_WALES	-1.012
SECONDAR	-1.440

In fact, a different picture emerges when the composite severity estimates are taken into account (as was also done for the Dutch secondary school teachers in the above). Since *BED*, *WINE* and *DEAD* all feature among the “upper” or “upper-intermediate” ranges in the three hierarchies of error, it may be difficult to claim that the lecturers had a tendency to overestimate the importance of the first two, but possible to argue that they may have slightly underestimated *DEAD*. In the GA version, *CAR* and *TO_WALES* are also included in the “upper” range of significant errors, which means that it is reasonable here to claim that the lecturers may have underrated their effects on GA judges (albeit only marginally so). This is different in the case of *SECONDAR*. As the latter has a composite severity estimate of 2.002 in the RP version and 1.821 in the GA form, these are neither significantly different nor particularly high in either version; this would make it hard to insist that its importance has been underestimated. At the same time, *OFF*, *THAT* and *ICE* do not actually feature in the upper or upper-intermediate ranges of any of the three hierarchies, so there does appear to be a certain basis for claiming that the lecturers may have somewhat overestimated their importance – especially as regards RP (see also 3.5.7). Finally, as the composite severity estimates for *HOT_TEA* and *NEW* are not particularly high in any version of the Native-speaker Experiment either, the Dutch lecturers cannot easily be faulted for underestimating their significance.

This implies that Dutch lecturers involved in pronunciation teaching would probably do well to attach more importance to the glottal replacement of lenis stops (*DEAD*), and possibly to the effects of *r*-deletion (*CAR*) and incorrect phrasal stressing (*TO_WALES*) – at least if their model is GA. They may also find it helpful to be aware of their slight tendency to overemphasise the significance of devoicing in *OFF*, *TH*-stopping in *THAT* and incorrect vowel length and quality in *ICE* if their pronunciation model is RP. In some cases, they may find their judgements of errors such as those illustrated by *BED*, *WINE*, *SECONDAR*, *HOT_TEA* and *NEW* to be at odds with those of particular groups of native speakers. Their priorities for the other ten tokens, however, do not appear to be strikingly different from native-speaker judges. This is, however, all on the assumption that the two experiments are sufficiently similar for there to be a basis for comparison.

Purely based on the number of outliers produced in regression analysis, it is striking that the Dutch lecturers judged more tokens differently from the native-speaker respondents than the Dutch secondary school teachers (and, as will be shown, than the Dutch students). The correlation coefficients are also a little lower (.627 for the RP version and .617 for the GA form), suggesting that the correlation between the lecturers and the two groups of native speakers is somewhat weaker.

One may be led to infer from this that the Dutch lecturers are a little less in tune with native-speaker pronunciation judgements than the secondary school teachers. It would also be interesting to hold this up against the notion that teachers tend to judge learners’ errors more severely than non-teachers (as in Hughes & Lascaratou 1982). If this is true, and if, as Davies (1983) suggests,

teachers' judgements are affected by the educational context in which they assess the errors, the present results would suggest that in this respect, some allowances will have to be made for differences between groups of teachers as well. Divergent levels of "linguistic sophistication" may be a factor here, especially since, according to Johansson (1978: 22), this "may be an obstacle rather than an advantage in judgements of acceptability". Bongaerts (1999: 9), however, found that native speakers who are experienced teachers or judges of pronunciation do not assess non-native pronunciation any differently from individuals who are not.

It should be noted that many of the lecturers that took part in the Dutch Experiment are not involved in pronunciation training, but instead teach literature, philology or linguistics. Conversely, all secondary school teachers will be involved, to a greater or less extent, in proficiency teaching, as this is the main focus of the English curriculum in secondary schools in the Netherlands (see also 2.2.1). It is, however, unclear to what extent *pronunciation* training features in this. Interestingly, only one out of the 97 secondary school teachers indicated that they paid no, or hardly any, attention to pronunciation in class, as opposed to no fewer than 20 out of 62 lecturers. Needless to say, the lecturers' relative lack of involvement in pronunciation training does not reflect on their "linguistic sophistication", but in some cases it may have affected certain of their judgements. Nonetheless, it should be pointed out that as many as 37 out of 95 Dutch university students of English stated they had received little or no pronunciation teaching in secondary school. If almost 40 percent indicate this, this may well reflect on the actual attention to pronunciation given by secondary school teachers in general. However, it does not necessarily relate to those participating in the experiment, who may have been motivated to take part because of their interest in pronunciation (see also 2.2.1).

The comparisons involving Dutch university students of English produce some interesting results (see Tables 3.91 to 3.94). On the one hand, the correlation between their judgements and those of the two native-speaker groups is only moderate (.609 for the RP version and .58 for the GA form). Based on the number of outliers, however, it can be said that Dutch students judged fewer tokens differently from the native-speaker respondents than did either the Dutch secondary school teachers or the Dutch lecturers. For advocates of pronunciation teaching, it would be tempting to attribute the latter to the success of the training many of these students have been subjected to, while opponents may wish to argue that students appear to be more capable of internalising native-speaker norms than those who are supposed to teach them these. In any event, the difference between students on the one hand and lecturers and teachers on the other does reflect the notion that the latter are stricter judges of learners' errors. Needless to say, it is actually very unclear if any such conclusions may be drawn from a comparison of such disparate experiments.

Table 3.91. Regression coefficients for NL/STU, predicted from the RP version (R = .609).

	Coefficient	Standard Error
A	0.404	0.986
B	1.051	0.306

Table 3.92. Outliers in the regression analysis of NL/STU and the RP version, using a criterion value of 1 Likert scale point for the standardised residual.

BED	1.286
THAT	1.405
ICE	1.139
HOT_TEA	-2.999

Table 3.93. Regression coefficients for NL/STU predicted from the GA version (R = .580).

	Coefficient	Standard Error
A	0.975	0.885
B	0.833	0.262

Table 3.94. Outliers in the regression analysis of NL/STU and the GA version, using a criterion value of 1 Likert scale point for the standardised residual.

THAT	1.156
RED	-1.255
HOT_TEA	-3.144

In addition, there are still a few tokens that the students attached either too much or too little importance to. Judging by the native speakers' AS estimates, these include THAT and HOT_TEA for both the GA and RP versions of the experiment. Especially the latter has a very low residual (-2.999 for the RP form and -3.144 for the GA version), which means that HOT_TEA was rated far lower by the Dutch students than would be expected from the general pattern established by RP and GA respondents' relevant estimates. Conversely, the Dutch students were slightly more strict on THAT than the correlation would lead one to expect. Dutch students also appeared to overestimate the importance of BED and ICE slightly, at least in comparison with the RP judges' AS estimates, and they had a weak tendency to underestimate the significance of RED, but only as compared with North American listeners.

If the composite severity estimates are factored in, however, it turns out that one error that the students seemed to have underestimated did not rank very highly in the relevant hierarchies (HOT_TEA), while another pronunciation issue that they appeared to have overrated is in fact one of the highest-ranking (BED). But since THAT and ICE were not included in the upper or upper-intermediate ranges of any of the three hierarchies of error, there is some basis for claiming that students may in fact have overestimated these errors. (Some of these students will be relieved to learn that TH-stopping in this high-frequency word is less salient to native speakers than has previously been assumed, especially where judges from the British Isles and the Antipodes are concerned; see also 3.5.6.) In addition, there is also some support for the notion that they slightly underrated uvular realisations of /r/ as compared with GA judges, as the error in RED is among the most salient in the GA hierarchy.

Significantly, this suggests that Dutch students would do well to be aware of North American objections to uvular realisations of /r/. In addition, they may find it useful to recognise their slight tendency to overemphasise a number of pronunciation issues, including the notorious problem of TH-stopping in high-frequency words such as *that* and, if their model is RP, the error exemplified by ICE (vowel length in /aɪ/). Finally, they should consider to what extent the importance they accord to degemination (as in HOT_TEA) and devoicing (BED) fits in with the priorities given to these by native speakers. As far as the other 17 tokens are concerned, students did not appear to judge them very differently from either group of native speakers – even though this may be of limited relevance in view of the dissimilarities between the two experiments.

It may be disappointing to learn that these comparisons of the pronunciation judgements of different groups of judges do not provide clear support for or against the notion that non-native speakers – and teachers in particular – may be stricter than native speakers when it comes to evaluating learners' errors, or favouring segmental to prosodic errors. What does emerge from these analyses, however, is that, in this respect, researchers should be careful to make distinctions between different groups of teachers, and differentiate between specific errors and error categories (especially as regards the learners' target accent). For instance, there are both phonemic and sub-phonemic errors among those that Dutch teachers, lecturers and students may have a tendency to underestimate. In any event, all parties concerned would do well not to underestimate the English pronunciation problems caused by the absence of aspiration and the incorrect use of glottal stops. Particularly those whose model is GA should be aware of the effects of r-deletion, uvular realisations of /r/ and also possibly those of phrasal stressing, and perhaps even ascribe a little less significance to TH-stopping in high-frequency words such as *that*. The latter would also be recommended to those whose model is RP, who in particular should also consider according less significance to devoicing in high-frequency words such as *of* and to Dutch realisations of /aɪ/. All concerned may in fact find it useful to distinguish between RP and GA when it comes to hierarchies of error (as has also been pointed out in 3.2.3). It would also be a good idea if teachers, students

and lecturers alike considered to what extent their attitude to phenomena such as degemination, dialect mixing, devoicing and dark [ɫ] is actually in keeping with those of the native speakers whose accents they may be using as a model.

3.7 General discussion and preliminary conclusions

A crucial presupposition in the preceding sections of this chapter is that it is the combined effects of detection and assessment that contribute to error severity, allowing errors to be ranked in a hierarchy and clustered with others of roughly equal importance. Admittedly, it is possible, and indeed desirable, to separate the effects of detection and assessment, since their different effects account for some of the variation between estimates. Nevertheless, it has been assumed that errors that had been assessed quite severely, but had been detected only rarely, should not be very highly ranked in any hierarchy – nor, at least in theory, should any errors that are widely reported but generally considered to be insignificant. Both these types of error would have a relatively low *composite severity*, and therefore only rank in the lower ranges of the hierarchy of error. Conversely, errors that were both frequently detected and also judged strictly would feature in the upper categories.

The estimation and analysis of the composite severity scores described in the preceding sections of this chapter shows that a hierarchy of error can be established with nine discrete error clusters. These range from representing incorrect stress (with estimates exceeding 3.5 Likert scale points) to those illustrating non-native intonation and certain distributional or realisational differences (which have estimates lower than one Likert scale point). The intermediate ranges consist of phonemic errors, together with the other distributional, realisational, and suprasegmental errors (the latter including a particularly high-ranking error that could also be interpreted as illustrating incorrect stressing).

It is interesting to note the degree of importance accorded to stress by all groups of speakers. In a similar experiment on British reactions to the L2 English of native speakers of Swedish, Johansson (1978: 106) already found that “one example of incorrect placement of stress” included in the test was considered a significant error. Even though Jenkins (2000: 150) considers “word stress” to be “something of a grey area”, virtually all other textbooks suggest that stress errors are salient to all native speakers. Clearly, these native-speaker judges do not need a clear context to identify and evaluate such errors. The absence of such a context, however, may help to explain why intonation and other suprasegmental errors were accorded a much lower severity in the present experiment – although other factors are also known to have affected similar experiments involving these phenomena. As a result, it is more difficult to assess the effect of such phenomena on pronunciation as “serious” than the writers of certain manuals and textbooks would suggest (see 3.5.23). In addition,

suprasegmental phenomena may well be particularly more salient upon repetition, or may be perceived by naive judges as segmental errors instead. While Dutch intonation, at least in the analyses of the present core experiment, does not appear to be a significant source of error, it may require more research into its effects on native speakers to establish this with greater confidence. Such unmistakable effects were, after all, found for Swedish (Johansson 1978: 111) and certain other languages (as discussed in Johansson 1978: 109, Anderson-Hsieh *et al.* 1992: 531–534, and Munro & Derwing 1995: 76) – although one wonders whether in the case of Swedish this may be ascribed, at least to some extent, to the well-known distinctive nature of Swedish intonation. As Anderson-Hsieh *et al.* (1992: 548) suggest, “the effects of the segmental ... error rates on pronunciation scores *relative* to prosody may be more dependent” on a speaker’s “native language” [their emphasis]. In any comparison of studies of error gravity which involve different target languages, the possibility of such intrinsic differences should be taken into consideration.

It would perhaps be expected (as may be true of some Dutch judges) that respondents would consider phonemic errors much more important than those of a realisational or distributional nature. After all, it is frequently claimed or assumed that the most realistic or desirable pronunciation target for non-native learners of English (or indeed any other natural language) is merely to get the message across and that learners should concentrate primarily on phoneme contrasts rather than on realisational or distributional errors that do not impede this process. Munro & Derwing (1995: 93) go so far as to state that “[i]f comprehensibility and intelligibility are accepted as the most important goals of instruction in pronunciation, then the degree to which a particular speaker’s speech is accented should be of minor importance”. Similarly, while Jenkins (2000: 159) allows “close approximations to core consonant sounds” in her *Lingua Franca Core*, she specifically excludes “certain approximations ... where there is a risk that they will be heard as a different consonant sound from that intended”. If, like Jenkins, “we are mainly concerned with intelligibility not for native speakers but for other L2 speakers of English” (Jenkins 2000: 158), the problem remains that in the *Animal Farm* of International English, some phoneme confluations are more intelligible than others! A Polish or Turkish listener may have fewer problems with Dutch final devoicing in English than a speaker of French, Farsi, Hindi or Hungarian. If, notwithstanding Jenkins, native speakers are still included in an English learner’s target audience, it is debatable whether avoidance of phoneme conflation should be the learner’s main concern.

In his analysis of his own experiments on English native-speaker reactions to Swedish pronunciation errors, Johansson (1975: 82) had already refuted the argument that “subphonemic deviations are of minor importance” and had shown this to be based on the untenable claims that they are neither perceived by native speakers nor impede communication. The latter, he states, is based on “a very narrow concept of communication”, which does not allow for the fact that such deviations “attract the attention of the listener and make him concentrate on the medium rather than the message” and may even make

speakers who are aware of their own deviations feel “maladjusted” and inhibited (1975: 83). In one of Johansson’s (1978: 127) own experiments, however, he also discovered a tendency (with only a few exceptions) for native speakers to prioritise phonemic errors. In any event, it should be noted that there are certain phonetic differences between realisations which, if unobserved, may cause native listeners not to perceive certain phoneme contrasts, such as glottalling or, as Jenkins also points out (2000: 159), aspiration of stops.

Even if the results of the Native-speaker Experiment appear to show an overall tendency for the phonemic errors in question to outweigh distributional or realisational ones, there are also clear counter-examples. Schwa-insertion in words like *film* and a uvular realisation of /r/, for instance, rank among the upper-intermediate errors, a result which is in keeping with the hierarchy of error presented in Collins & Mees (2003b: 291). As respondents’ comments showed, here it is the stigma attached to such pronunciations that may cause irritation or amusement – because they are associated either with regional or social accents, or with caricatures of foreign speech. This is in accordance with Johansson’s (1975: 32) findings that the degree of irritation created by an error plays an important role in establishing its “gravity”. (For a different view of the role of irritation in error gravity, see Albrechtsen *et al.* 1980: 395.) At the same time, native-speaker judges in the present experiment considered classic phonemic errors such as TH-stopping or devoicing to be much less important – at least in high-frequency function words such as *that* or *of*. Conversely, the considerable significance given in the present experiment to glottalling of lenis stops can only be explained if this is seen as an error which causes phoneme conflation.

It is noteworthy that while some L2 English pronunciation errors by Dutch native speakers are perceived as markedly foreign, others (or sometimes even the same errors) are associated with regional or social variation in native English. The latter tendency may have been compounded by the presentation of errors, in the Native-speaker Experiment, as single deviations in otherwise natively pronounced sentences.⁹ In much foreign-accented English, however, there is usually a layering of mistakes rather than one single unexpected feature (see Abbott 1991). Collins (1979b) refers to this phenomenon as “a mosaic of errors of varying degrees of gravity layered one upon another”. This effect may even have caused native speakers to underestimate the importance of certain errors which are unlikely to cause unintelligibility in isolation, especially as the judges were presented with the spelling of the entire sentence as they heard the audio recordings. As Prator (1968: 19) points out, “a language teacher would be well advised to regard unintelligibility not as the result of phonemic substitution, but as the cumulative effect of many little departures from the phonetic norms” – an important observation also adduced by Koster & Koet (1993: 90). Nevertheless, it is useful for Dutch learners to be aware of the fact that elements of their pronunciation may be perceived by native speakers as being within the

⁹ A similar effect has also been observed by Norell (1991: 66).

range of indexical variation in native English. Furthermore, this should also serve as a warning to L2 learners of English with different L1 backgrounds.

Whether or not native speakers' association of such single errors with regional or social variation in English will cause them to be more lenient is another matter. Respondents' comments suggest that this may not always be the case. There is even some indication that some pronunciations associated with Ireland or Scotland (see 4.6) were not judged more leniently by respondents from those areas – a subject that will be taken up in more detail in Chapter 4. In addition, pronunciations that are stigmatised in one part of the English-speaking world may be seen as harmless (or even “charming”) regional or ethnic markers in another; for instance, this appears to have affected judgements of TH-stopping in *that*. While the varieties in Britain that are associated with this phenomenon, such as Afro-Caribbean English and Irish English, may not be among the most prestigious (especially the former), the American varieties that have TH-stopping include AAVE, which, according to Milroy (1994: 189), is much more heavily stigmatised than either of the two varieties used in Britain. Such results do not lend any support to commonly made claims that because a particular realisation is found in native varieties of English, its use by non-native speakers is likely to be uncontroversial in all contexts. For instance, it is not consistent with Jenkins's assertion (2000: 27) that it is “no longer appropriate to regard ... variation from the L1 as automatically deviant” since “[m]uch of it comprises acceptable regional variation on a par with that which we find among L1 accents of English”.

If teachers of spoken English would like learners to communicate effectively with native speakers (and others), they would do well to include a certain level of phonetic detail in their pronunciation training, not only to ensure intelligibility but also to obviate listener fatigue or any attitudinally marked responses – which can be very emotive, as respondents' comments demonstrate. The importance of phonetic detail appears to be a crucial assumption underlying pronunciation manuals aimed at Dutch-speaking learners such as Gussenhoven & Broeders (1997) and Collins & Mees (2003b), and is borne out by the analysis of the two experiments. Advanced learners in particular will also require some of the sociolinguistic competence of a native speaker of their intended target accent in order to judge the acceptability of their own (and others') deviations from the model, especially if they are sensitive to the kind of native-speaker stigmatisation apparent from some of the comments. On a more positive note, awareness of such sociolinguistic patterns may also help to encourage learners' efforts to acquire an authentic and localised variety of English, rather than some elusive “international” model. This is especially relevant to learners with integrative motivations, which, as Ellis (1994: 513) concludes, “has been shown to be strongly related to L2 achievement”. All this suggests that pronunciation training would benefit from *more* rather than *less* attention to phonetic and sociolinguistic detail.

Such training would appear to be relevant to communication with native and non-native speakers alike. As the many similarities between the Native-

speaker Experiment and that involving Dutch respondents unmistakably demonstrate, EFL learners will tend to judge English pronunciation errors not simply by the standards of their own native language (i.e. *endonormatively*), but by those of what has been called their “interlanguage” (Selinker 1972) or “approximative system” (Nemser 1971). This interlanguage (IL) is likely to include, amongst other things, a degree of approximation, no matter how distant or stereotyped, of native-speaker standards.¹⁰ Ellis (1994: 213) also notes that in “foreign-language settings the preference model is nearly always a standard variety of the inner circle”. In particular, EFL learners interested in achieving “additive bilingualism” (as opposed to using more “restricted” varieties of international English) are likely to refer to such *exonormative* standards.¹¹ There is little evidence to suggest that such learners are a “diminishing breed”, as Jenkins (2000: 220) claims. Bruthiaux (2003: 168) has pointed out that in spite of the “well-documented claim for variety status to be accorded to English used as a lingua franca (or “ELF”) by second language users among themselves ... the domain of such language use remains restricted to specialized transactions (business negotiations, industrial cooperation, tourism, etc.) by a relatively small number of speakers, and broader variety-creating conditions remain largely absent”.

In fact, student complaints about lecturers’ foreign accents in English (e.g. Klaassen 2002) anecdotally suggest that the notion of native-speaker standards is not necessarily found in very advanced or fluent learners only. There are also studies which show the considerable biases EFL learners may have against non-native English (as discussed in Major *et al.* 2005), which it would be patronising to dismiss as mere “linguistic insecurity”, a phrase used by Jenkins (2000: 211–212). In any event, advanced learners are unlikely to abandon all reference to *exonormative* standards when communicating in English with other non-natives, especially since this may provide these learners with a more useful context for understanding their interlocutors’ English than any *endonormative* considerations (unless perhaps these people speak a closely related language).¹² An increased sensitivity to phonetic detail may in fact increase their tolerance and understanding of unfamiliar varieties of English.

¹⁰ According to Tarone (1988: 43), however, the “interlanguage norm ... may sometimes contain accurate target-language variants, but may as often contain prestige native language variants or even uniquely IL prestige forms”.

¹¹ See Ellis (1994: 208, 221) for a discussion of the terms “additive bilingualism” and “restricted” variety, and Melchers & Shaw (2003: 32) for a discussion of the terms “endonormative” and “exonormative”.

¹² This is in spite of Kachru’s unprovable claim (quoted in Ellis 1994: 221) that in interactions between non-native speakers “the *British* English or *American* English conventions of language use are not only not relevant, but may even be considered inappropriate by interlocutors. The culture bound localized strategies of, for example, politeness, persuasion and phatic communion ‘transcreated’ in English are more effective and culturally significant”. In this context, it may be useful to refer to an experiment by Bansal (1965/66, 1969), described in Johansson (1978: 9), which showed that Nigerian and German respondents had

In spite of what the supporters of English as an International Language may claim, it is an expansive rather than a reductionist approach to pronunciation teaching that will help learners adjust the exonormative standards they may have partially internalised to an approximation of actual native-speaker norms sufficiently close to enable them to meet their communicative needs. This is likely to benefit their communication with both native and non-native speakers. In addition, awareness of phonetic detail and sociolinguistic competence are necessary to appreciate the differences between groups of native speakers when it comes to judging and ranking error severity. The results of the Native-speaker Experiment show that if composite severity estimates are ranked separately for the two main groups of respondents (British, Irish and Antipodean judges versus North Americans), the resulting hierarchies of error are considerably different. There is additional variation *within* these two main groups, although this is statistically significant only for a few tokens. All this makes it more difficult to conceive of native-speaker pronunciation norms as a single and immutable system that can, in the interest of didactic or other considerations, easily be reduced to a common core. It also argues in favour of broadening learners' sociolinguistic competence. While this would be relevant to all learners wishing to acquire a pronunciation that is acceptable to different groups of native speakers, it is essential for those who aspire to a near-native level. This is also advocated by Bayley & Regan (2004: 325), who state that "far from being a peripheral element, knowledge of variation is part of speaker competence". This suggests, in their view, that "second language learners also need to acquire native-speaker ... patterns of variation" (2004: 325).

Overall, the hierarchies for the RP and GA versions are similar in that the tokens illustrating stress and phonemic errors tend to be found in the higher ranges and those exemplifying distributional, realisational and suprasegmental errors in the lower (with a number of counterexamples). In addition, four tokens illustrating phoneme confluents (including /æ ~ e, w ~ v/ and two examples of TH-stopping) and two others (representing stress and schwa epenthesis respectively) were considered to be similarly serious, while degemination, the trans-Atlantic version-external realisation of *secondary* and the two intonation tokens were judged to be equally unimportant. This has two implications. Firstly, there is uniformity of judgement when it comes to a small number of errors, interestingly enough including one item (SECONDAR) that some respon-

greater difficulty understanding Indian English than "British, American or Indian listeners". Similarly, Major *et al.* (2002: 173) found that "both native and nonnative listeners scored significantly lower on listening comprehension tests when they listened to nonnative speakers of English" than when listening to native speakers. In addition, Major *et al.* (2005: 63) also found that "ESL listeners experienced more difficulty with the ethnic and international dialects of English [represented in their study by AAVE, Indian English and Australian English] than with Standard American English, but not with the regional dialect of English". This goes against Jenkins's (2000: 206) claim that "the assumption that a 'standard' 'N[ative] S[peaker]' accent is internationally intelligible is a myth".

dents could be expected to object to as being iconic of the other major variety of English, but did not. Secondly, the large majority of tokens (22 out of 32) were judged significantly differently in the two versions – the distractor included. Amongst these were two other examples of “token mirroring” (**r**-deletion/retention in *car* and yod-insertion/deletion in *new*) that were indeed assessed more severely in one of the two versions, possibly partly as a result of local stigmatisation (which is non-existent in the case of SECONDAR). North Americans’ somewhat higher composite severity for CAR (with significantly higher hit rates) suggests that **r**-retention where the prestige variety is non-rhotic is slightly less severe than **r**-deletion where the prestige variety is rhotic. Remarkably, the fact that the overwhelmingly rhotic Irish respondents did not judge this token significantly differently from the exclusively non-rhotic RP speakers implies that, at least in this instance, the former are quite capable of assessing Dutch pronunciation by non-Irish (i.e. *exonormative*) standards. Furthermore, while CAR was not viewed as particularly serious in either version of the experiment, the RP judges’ composite severity for NEW exceeded those of GA judges by more than one point on the Likert scale. The same is true of a comparison between self-identified speakers of these varieties. Clearly, it is difficult to predict which pronunciation features are important to which group from a simple comparison of the differences between their accents. As Johansson (1978: 31, n. 5) points out, “there is no necessary correspondence between linguistic measures and communicative efficiency”. It would seem that sociolinguistic competence is required to make these finer distinctions.

Clearly, the different priorities given to particular errors by judges of these varieties cannot merely be predicted by the features that distinguish RP from GA – as is also evident from the differences in significance being accorded to phonemic errors involving either vowels or consonants. Table 3.95 shows that, while in the RP version the tokens representing phonemic vowel contrasts all have high severity estimates of roughly around 3.0 on the Likert scale, the same errors are ranked very differently in the GA version, with STOOD and COLOUR ranking far below BAT and also below any of the errors involving consonants. While the dramatically different estimates for STOOD and COLOUR in the GA version are the result of much lower hit rates, this is not something that may be predicted from a comparison of the phoneme inventories of GA or RP. In fact, previous research (Johansson 1978: 97, 111, Koster & Koet 1993: 77, Munro & Derwing 1995: 76) would seem to suggest that native speakers of English (unlike native speakers of Spanish, see Schairer 1992) attach more significance to consonant than vowel errors, so one would expect to see this reflected in all three versions. This is clearly not the case. Even if the ranking in all three versions has been affected, at least to some extent, by this particular selection of vowel errors, there still appear to be different trends in the two versions when it comes to prioritising consonants to vowels.

Table 3.95. Ranking of severity estimates for tokens representing phonemic errors, by version (overall, RP and GA). Cells shaded grey represent phonemic errors involving vowel contrasts; the remaining errors and estimates represent consonant contrasts.

Overall		RP version		GA version	
Token	Estimate	Token	Estimate	Token	Estimate
THIN	3.416	THIN	3.468	BED	3.660
AUTHOR	3.204	AUTHOR	3.250	VAN	3.546
VAN	3.117	COLOUR	3.125	AUTHOR	3.305
BED	3.057	BED	3.018	THIN	3.233
BAT	2.958	BAT	3.004	WEATHER	3.225
WINE	2.745	VAN	2.938	BOTH	3.151
COLOUR	2.740	STOOD	2.918	BAT	3.088
WEATHER	2.480	WINE	2.713	WINE	2.828
STOOD	2.317	BREATHE	2.265	BREATHE	2.513
BOTH	2.315	WEATHER	1.642	THAT	1.817
BREATHE	2.280	BOTH	1.582	OFF	1.382
THAT	1.196	OFF	0.981	STOOD	1.004
OFF	1.115	THAT	0.929	COLOUR	0.995

According to Abbott (1986: 228), the idea that English vowels are less important for general intelligibility, and should therefore be less of a priority in pronunciation teaching, is in fact a fallacy. Abbott (1986: 229) argues that it is based on dubious attempts to put to “serious pedagogical use” the observation that replacing all English vowels by schwa may still result in intelligible speech (see also Johansson 1978: 97). In any event, the results of the two versions do not show a common trend that allows for any conclusions to be drawn about the relative insignificance of vowel errors. What does emerge, however, as has been indicated before, is that there are certain specific errors that are much less of a priority in one major English pronunciation model than the other, which cannot be predicted from a comparison of these (although regional variation in US English may in fact have affected the severity estimate for COLOUR). In addition, these differences are not the result of disparate evaluations, but of different rates of detection. While those North Americans who detected these errors judged them as severely as the other native speakers, this should be weighed against the fact that these judges were relatively small in number. This suggests that there may be groups of native speakers with very strict and possibly “idealised” pronunciation norms for errors that are hardly detected by the vast majority of judges with similar accents.

Differences in detection rates are also an important factor in accounting for inter-version differences in the evaluation of the other tokens. Some of these

differences may be ascribed to the performance of the two actors (notably as regards BOTH and WOULD_ON). Nevertheless, it is highly probable that a number of typically Dutch errors which may have the effect of (1) neutralising the fortis/lenis contrast (BED, VAN, OFF, DEAD), of (2) TH-stopping (BOTH, THAT, WEATHER) or (3) of consonant deletion (CAR, TELL) were detected more readily and/or assessed more severely by North Americans than by other respondents *not* because they are unintelligible but because of other reasons. They may, for instance, have been associated with stigmatised native speech or possibly with caricatures of certain foreign accents (such as a uvular realisation of /r/ in RED). This would appear to be important information for anyone in the Netherlands or Dutch-speaking Belgium teaching or learning American English. It may even be seen as a warning to those who are interested in learning any of the varieties which have these features (such as AAVE and non-rhotic regional varieties) and are inclined to incorporate these elements in their speech. As noted in 3.6, teachers and students of English should be aware of the different attitudes between judges of RP and GA in these and other respects, and consider if their own attitudes to dialect mixing and certain features associated with stigmatised speech are in fact consistent with those held by native-speaker judges of the target accent in question.

Remarkably, North Americans judged only three of the six examples of TH-stopping more severely than did their British, Irish and Antipodean counterparts (BOTH, THAT, WEATHER), and even attached slightly less importance to one such token (THIN). Although both versions show a tendency for substitutions for /θ/ to be ranked higher than substitutions for /ð/, in the RP version BOTH is actually accorded the lowest importance of all errors involving /θ/ or /ð/ – possibly due to the actor's performance – whereas in the GA form WEATHER does not rank significantly lower than THIN, AUTHOR or BOTH, which could be a result of some GA judges perceiving a Dutch realisation of *weather* as *wetter*. For all versions, it is true to say that, of the errors involving /ð/, WEATHER and BREATHE rank significantly higher than THAT (which features the high-frequency grammar word *that*). While the influence of contextual factors (such as an actor's performance, word frequency or the absence or presence of a minimal pair) may attest to some of the experiment's limitations, the varied responses to TH-stopping in initial, medial or final position still suggest a useful addition to previous research into error gravity. A similar case may be made for the relative importance of /f ~ v/ confusion in initial (VAN) or final (OFF) position.

In spite of the experiment's limitations, the results nevertheless support the intuitive notion that certain errors may be more or less salient depending on their position in the word. This may vary for different groups of native-speaker judges. This is not reflected in existing pronunciation textbooks and studies of error gravity, which tend to discuss error location within a word only in the context of errors that typically occur in a particular position, such as the consonants affected by final devoicing. It would, however, be useful for Dutch teachers and learners of English to be more aware of how the position of an

error in a word may increase or decrease its chances of being detected, and adjust their priorities accordingly. On a practical level, this would mean, for instance, that in a sentence such as “Think twice before you let them get together” (Collins *et al.* 2001: 23), Dutch learners would be encouraged to pay more attention to avoiding stop realisations of /ð/ in *together* than of the same phoneme in *them*. Similarly, inter-version comparison of the estimates suggest that aspiration should be accorded less importance in pronunciation training to learners of GA. In fact, so few North Americans detected the error in TIE that one wonders if initial aspiration of stops is indeed such an important acoustic cue for GA.

While it is striking that so many tokens were judged differently in the two versions, it is perhaps even more remarkable that the listeners in the two groups also showed a more general tendency to detect and evaluate errors differently. Admittedly, the two groups’ overall composite severity does not differ demonstrably, but if this is broken down into HR (hit rate) and AS (adjusted severity) estimates, it turns out that RP listeners tend to have higher HR and lower AS estimates than their North American counterparts. In fact, GA listeners’ AS estimates were *invariably* higher. This shows that RP listeners notice a great many deviations which they are not prepared to classify as serious errors, while GA listeners tend to consider most detected errors as serious. More research is needed to establish if these results reflect fundamentally different attitudes to Dutch-accented English as being either “noticeable but not serious” or “serious only where noticeable”.

In any case, if North Americans object more to clearly identifiable errors, this may be taken to mean that they are inclined to admit to less tolerance of non-standard and non-native speech than the British, Irish and Antipodean groups combined. Surprisingly, while this goes against the Dutch stereotype that British and Irish native speakers are stricter judges of Dutch English pronunciation vis-à-vis Americans and Canadians (see 1.1), it does not necessarily imply that the latter group are more “ethnocentric”. It may be true, as Milroy (1994: 178) states, that in the US, foreign accents “seem to be more subject to negative evaluation than in Britain”, but, as was argued in 3.4.4, tolerance of accented speech may also be inspired by covert motivations that are exclusionist rather than integrative. As Corder (1973: 61) points out, native speakers assign a “special role” to foreigners “in which behaviour, inappropriate in a native, is socially acceptable” (see also Johansson 1978: 128–129, Ellis 1994: 213, Scheuer 2005: 112).

Of course, a native speaker may have other reasons to be lenient: for instance, Nickel (1972: 19–20) refers to “ein Gemisch von Dankbarkeit und Stolz, daß seine Muttersprache hier Lehrgegenstand ist” (see also Johansson 1978: 119). One of many other factors is awareness of linguistic variation. Nickel (1972: 20) even suggests that in the case of the native speaker, “die Breite seiner Sprachkenntnisse sowohl im Englischen als auch im Amerikanischen stimmen ihn insgesamt großzügiger”, although this appears to be different for British and American judges. In extreme cases, completely

monolingual speakers may even express their leniency towards foreign-language errors implicitly to stress the inconsequentiality of second-language learning at an advanced level, and thus to reaffirm the monolingual norm. Of course, such a tendency to “damn with faint praise” need not have informed comments by those respondents who find pronunciation errors “charming” or a handy detection mechanism for foreign speech. Nevertheless, learners would be well advised not always to take statements such as “I really like your accent” at face value. If, however, it is one of the goals of EFL teaching to empower learners to deal with the realities of interaction in a second-language environment, teachers would do their students a disservice if, when considering their teaching priorities, they did not take both positive and negative attitudes to foreign accents into consideration.

Although it would be useful to Dutch learners and teachers of American English to be aware of North American attitudes to “overt” foreign or stigmatised pronunciations, they should also realise that Americans and Canadians tend not to detect or report all Dutch errors as readily as some other groups do. A similar effect was found for female judges in general. Even though women’s overall composite severity is lower than men’s, their AS estimates are invariably higher than their male counterparts’, who mostly had significantly higher HR estimates. This appears to be consistent with what Labov (2001: 266) has termed the “general linguistic conformity of women”. While older respondents’ lower detection scores may be ascribed to reduced auditory sensitivity (Sommers 2005), older judges’ generally lower adjusted severity may perhaps be seen in the context of their greater exposure to language variation (see also Major *et al.* 2005: 45). Judges who identified themselves as strict also tended (with a few interesting exceptions) to have higher HR and AS estimates.

One may be led to infer from this that younger men from the British Isles or the Antipodes would detect pronunciation errors most readily, and younger women from North America would assess pre-identified errors most severely, especially if they identified themselves as relatively strict. However, it must be remembered that the overall composite severity estimate (which combines detection with assessment) is in fact a little lower for both women and older respondents (but not for North Americans). This means that, based on overall performance, it is difficult to provide a clear profile of who in, for instance, a classroom situation, would be the strictest native-speaker judge.

Though not as pervasive as the contrasts between the RP and GA versions, there were still a number of striking differences between the various other accent groups. To start with, there was a tendency for judges from Scotland to identify themselves as being more lenient, and an opposing tendency for listeners from the American East Coast to categorise themselves as being stricter. These respondents’ self-assessments appeared to be justified, in that Scottish informants were indeed inclined to evaluate errors more tolerantly, while judges from the American East Coast actually tended to be more severe. A similar inclination towards tolerance was observed in judges from Australia, New Zealand and South Africa. A discussion of these results in terms of

ethnocentricity would not be justified – although it may be interesting to note that, according to a marketing survey held among European consumers (Steenkamp 1993), Scottish consumers were less ethnocentric than their English counterparts in buying consumer items from outside the UK. Steenkamp (1993: 22) attributes this to a Scottish tendency to identify the UK with England rather than Scotland. Be that as it may, the lack of enthusiasm in Scotland for an “English” accent such as RP (McClure 1994: 80, quoting the findings of Romaine 1980) may have prompted a few reservations about very strict evaluations against what is felt to be an exonormative standard (see Melchers & Shaw 2003: 32). A comparable effect was found to have influenced judges from Edinburgh in an experiment by Johansson (1975: 74). It may be hypothesised that a similar reluctance to judge Dutch pronunciation errors against RP influenced the evaluations of Australians, New Zealanders and South Africans – and possibly even affected their willingness to take part in this experiment in large numbers. Conversely, judges who identified themselves as hailing from the culturally dominant American East Coast may have a stronger emotional investment in GA than many other groups. All the same, it must be stressed that the above differences in severity are not very pronounced. For instance, the more lenient assessment of BAT on the part of the Antipodean judges is only evident from a post-hoc comparison of the AU&NZ&SA judges as against all the other groups combined.

Any expectations that Irish respondents would have reservations about judging Dutch pronunciation errors against exonormative RP standards were not borne out by the results. In two of the three instances where the Irish respondents’ composite severity estimates were significantly different from any of the other major accent groups in the RP version, these turned out to be higher (BED and VAN); in the third instance (FILM) the IRL respondents’ composite severity estimate was admittedly lower but their AS estimate was higher. This token also gave rise to quite a few comments from Irish judges which revealed high levels of stigmatisation. Remarkably, the incidence of negative comments in the IRL group is significantly higher than that of all other groups combined, implying that these judges may be more critical or more vocal as a group. Similarly to some North American groups, whose attitudes to certain potential pronunciation errors (such as BED and NEW) they would appear to share, IRL judges may be slightly less resistant to criticising foreign or stigmatised pronunciations in a linguistic experiment than other groups in the RP version. The exact cause of this effect is not clear.¹³

¹³ If the view is adopted that stricter evaluations of foreign speech correlate with higher levels of cultural homogeneity among judges, this does not help to explain the attitudes found among IRL respondents – if only because judges hailed from both the Republic and Northern Ireland. In addition, it should be noted that the post-1995 influx of immigrants into the Republic has contributed considerably to the country’s cultural diversity (Mac Éinrí 2001).

What is actually striking is that, apart from three tokens, the IRL estimates are not significantly dissimilar from those in the other RP groups. Admittedly, these results might have been different if more judges from the Republic and Northern Ireland had been willing to take part in the experiment, which it would be convenient to attribute, at least partly, to a possible antipathy to the actor's RP accent. In fact, it would be interesting to replicate the experiment using an Irish English-speaking actor instead (as well as a number of other "guises"). Based on the present results, however, the divergence is only slight, and does not even affect tokens such *THAT* or *CAR*, where significant differences from a group such as GB/RP may have been expected. In simple terms, this could either mean that both GB/RP and IRL listeners referred to the same common standard, or that one group (or both) incorporated their awareness of linguistic variation in their judgements. Comments indicate that there were both GB/RP and IRL respondents who adopted the latter approach. In the case of the Irish listeners, this would imply a certain willingness to adopt exonormative standards when judging foreign accents. A similar tendency can also be observed with the GB/NRP group and, to a slightly lesser extent, even with the Scottish and Antipodean listeners, whose severity estimates were only slightly more lenient. Not only does this suggest that Wells's (1982: 279) observation that "[e]veryone in Britain has a mental image of RP" should be extended to include other English-speaking countries, but it may also serve as a warning to EFL learners that certain native-speaker judges may evaluate foreign accents by the standards of a variety which they themselves do not use. Sometimes speakers of local or regional accents will even reject attempts on the part of EFL students to learn a variety other than the supra-regional standard (for instance as a form of "linguistic gatekeeping" or as a result of linguistic insecurity).¹⁴ This subject will be taken up again in Chapter 4.

While there were no significant differences between major accent groups in the GA version – which once again implies an inclination for the US/NGA and CDN groups to adopt an exonormative approach – some of these groups did in fact assess certain tokens differently from groups in the RP version. This may be taken to indicate that Canadians assessed *ICE* a little less strictly than did their American counterparts. While there is no conclusive evidence for this, it is likely that Canadians' well-known tendency to raise the diphthong in *ice* to [əi] has made them less prone to detect or reject other realisations of /aɪ/ that deviate from the GA standard. In view of the great many similarities between GA and mainstream Canadian English (Wells 1982: 491), it is perhaps unsurprising that there were no statistically significant differences between Canadians and the two American groups. While it is in keeping with the well-known stereotype of Canadian politeness that this group did not volunteer a single negative comment, this did not actually appear to affect their relative strictness (in terms of adjusted severity) as compared with respondents from outside North America.

¹⁴ For other examples of "linguistic gatekeeping", see Pavlenko (2002: 287).

Pairwise comparisons of any significant differences between self-identified speakers of RP, GA, regional varieties of British English and regional varieties of American English showed no striking deviations from the differences between judges in the RP and GA forms respectively. There was also an inconclusive trend for self-assessed speakers of standard British English to evaluate DEAD somewhat less severely, and THAT_THA more so, than speakers of regional varieties of British English, who, in turn, may have prioritised RED a little less (possibly as a result of the sporadic incidence of uvular-*r* in parts of North-eastern England). In addition, those who did not categorise themselves as speaking a standard variety of American English appeared to object even more to certain foreignisms or stigmatised realisations. This could be ascribed to greater linguistic insecurity, but possibly also to an increased awareness of the social consequences of using stigmatised speech. In any event, one serious limitation of the experiment is that respondents' self-identification as speakers of "standard" or "non-standard" English cannot easily be checked.

The relative scarcity of significant differences between major accent groups that cannot be explained in the context of inter-version variation suggests that one of the most important variables affecting respondents' judgements is the version of the experiment. This makes it possible to conclude that all those interested in learning or teaching English pronunciation should be careful not to confuse the effects Dutch pronunciation errors may have in the two varieties, or the different ways in which they may be detected and evaluated. The fact that most variation was attested between versions can also be seen as an actual limitation of the core experiment, since at least some of this variation may be derived from differences in performance between the two actors. In this light, it would be interesting to replicate the experiment using a number of additional "guises" such as Irish, Scottish, Canadian or Australian English, which are then also presented to native-speaker judges of English world varieties, as well as to Dutch learners and teachers of English. It should be remembered, however, that this would require using additional speakers, whose performance is unlikely to be identical to that of either the RP or the GA actor. It is extremely unlikely that one single actor can be found who can produce a wide range of convincing-sounding Dutch pronunciation errors in more than one native-speaker guise.

Replicating the core experiment with different groups of respondents from the Netherlands may throw new light on the analysis of the differences between natives and non-natives in 3.6. This would help to provide clearer support for or against the argument that some of these judges are stricter than others when it comes to evaluating the English pronunciation of Dutch learners, and that they also prioritise different types of errors over others. Be that as it may, the results discussed in 3.6 already indicate that Dutch teachers, lecturers and students may have a tendency to underestimate a number of both phonemic and subphonemic errors, such as the absence of aspiration and the incorrect use of glottal stops. They should also be aware of how different attitudes to foreign speech and to sociolinguistic variation may result in evaluations of Dutch pronunciation errors that can be strikingly dissimilar in, for instance, Britain or the US (see 3.6 for

details). Clearly, native-speaker norms are not monolithic but are subject to variation across time and space. In some cases, teachers and learners may find that they even overestimate the importance of certain errors, although they should be aware that native-speaker leniency may be inspired by factors other than the wish to accommodate non-native learners.

For all its limitations, the core experiment has produced a number of results that may be useful in realigning the error hierarchies that are directly or indirectly relevant to Dutch students of English pronunciation. For instance, while there has been a growing awareness among phonologists and phoneticians that perceptual salience is affected by syllable position (see Beckman 1999: 3, 20, Kingston 1985, 1990, Steriade 1993), none of the pronunciation manuals and textbooks consulted differentiate between the relative severity of TH-stopping and /f ~ v/ confusion depending on the position of the error in the word, with the notable exception of Collins & Mees (2003b: 286, 290). This makes it more difficult to interpret any claims (as in Koster & Koet 1993: 80) that native speakers of English attach more importance to /ð ~ d/ confusion than do Dutch judges. Similarly, both the stigma attached to TH-stopping and the special significance of /θ/-substitutions tend to be underestimated, as this cannot be totally predicted from functional load (Brown 1988: 222) or from considerations of learner difficulty (Jenkins 2000: 138). However, Dretzke's (1985: 149, 203) conclusion that /θ/-substitutions outrank those involving /ð/ is consistent with a similar tendency in the core experiment's overall hierarchy of error.

Since the explicit hierarchy of error proposed for Dutch learners of RP by Collins & Mees (2003b: 290) formed the basis for error selection in the two experiments, it is relatively easy to compare their ranking of selected errors with those evaluated in the core experiment. There appear to be few differences, with the notable exception of ICE (although it should be pointed out that the significance of this error may have been obscured by the design of the experiment). In spite of the high level of agreement, the results of the present experiment also imply that Dutch dark [ɫ] is less significant than the authors suggest (2003b: 291). In addition, it may be argued that the error analysis provided by Collins & Mees (1993: 124–130) for the benefit of learners of GA overstates the significance of errors such as TIE, COLOUR and STOOD. These errors are, however, quite significant to judges of RP. It is more difficult to compare the results of the core experiment with the "Hints for the future teacher" in Gussenhoven & Broeders (1997: 16–17), since they do not explicitly refer to many of the errors included in the test, or provide a clear hierarchy of error. The same is true of Koster & Koet (1993: 90), although their conclusion that Dutch teachers should continue to pay little attention to suprasegmentals may need to be re-examined.

The other studies of error gravity discussed in 3.5 are not directly concerned with Dutch learners' pronunciations, but it may still be interesting to compare their conclusions with the present results (to the extent that they are compatible). For instance, Dretzke's (1985: 203) hierarchy of error for German learners of English appears to underrate the importance of devoicing, /æ ~ e/ confusion and incorrect stress, and does not include errors such as STOOD and

COLOUR. Similarly, while Brown's (1988) attempt to relate error gravity to functional load is generally in keeping with the present results (for example as regards BED, BAT, VAN, WINE and COLOUR), his approach does not account for the significance attached to STOOD in the RP version (or the relative severity of /θ/-substitutions).

More significantly, a number of errors which are very salient to native speakers are spectacularly absent from Jenkins's Lingua Franca Core, such as those involving /θ, ð/ and certain suprasegmental features. In addition, many of her recommendations (such as "close approximations to core consonant sounds generally permissible" or "L2 regional qualities permissible if consistent") are insufficiently specific either to be compared to the present results or to serve as realistic guidelines for Dutch learners of English (Jenkins 2000: 141). Even though Jenkins's "Lingua Franca Core" is mainly intended for learners communicating with other non-native speakers, it would seem to be irresponsible to advocate a hierarchy in pronunciation teaching which pointedly ignores a number of native-speaker concerns. Arguably, the communicative effect on native speakers of stigmatised pronunciations cannot simply be removed by impractical proposals to inculcate greater sensitivity to non-native speech or to teach non-native realisations such as /ð/ and /θ/ substitutions to native speakers, as Jenkins (2000: 228) actually suggests. In the face of the world-wide sociolinguistic dominance of native speakers of English over non-native speakers, it would be better to provide learners with the tools they need to communicate effectively with native and non-native speakers alike. This is not a "conservative" attitude to language variation, as Melchers & Shaw (2003: 30) somewhat tendentiously claim, but a way of empowering those actually motivated to *acquire* a foreign language beyond the basics. It is suggested that in terms of the second-language curriculum, this implies rather more attention to phonetics and sociolinguistic competence, or, as Bruthiaux (2003: 175) put it (in a different context): "less liberation and more linguistics".

CHAPTER 4

ACCENT SIMILARITY

4.1 Token similarity to Dutch English

One of the principal objectives of the core experiment of this dissertation is to discover whether particular Dutch English realisations are judged more leniently by respondents who, on the basis of their accent, can reasonably be expected to be familiar with such pronunciation features, or even use these in their own varieties. If, for instance, respondents identify themselves as speakers of Irish English, they are likely to be quite familiar with the characteristically Dutch English feature of schwa epenthesis in *FILM*. Since this is also a well-attested feature of Irish English, some of these Irish respondents may well say [ˈfɪləm] themselves. As a result of this, they may be less inclined to consider this particular Dutch English realisation a serious error. If, however, the realisation is a familiar one but heavily stigmatised, as would appear to be the case with *FILM* in North America, respondents will tend to be stricter. Needless to say, if respondents are unfamiliar with particular realisations, it will be assumed that they will not be predisposed to them either more or less favourably.

If the majority of speakers in a particular accent group use a characteristically Dutch English pronunciation feature, they are more likely to recognise it as a familiar characteristic of their own variety rather than as a foreign or stigmatised feature. Even though it raises some interesting issues, this situation is relatively rare. In this experiment, only *BAT*, *TIE*, *CAR* and *STOOD* can be ascribed, with any degree of confidence, to a majority in one or more accent groups. Accent descriptions do not, as a rule, provide precise estimates of numbers of speakers using a particular realisation, and it would seem prudent to associate any such realisations with a minority unless there is clear evidence to the contrary. According to such a conservative estimate, there are as many as 20 tokens in this experiment that may be considered minority realisations in one or more minor accent groups. These 20 tokens include *BAT*, *TIE*, *CAR* and *STOOD* (which are minority realisations in some accents and majority realisations in others). Minority realisations are less likely to be familiar to all speakers of that variety, and may well be considered to be more foreign or more stigmatised. This is likely to affect the way respondents judge the severity of such features.

As has been stated in 2.1.3, 2.3 and 3.2.6, a number of tokens are not completely identical in the two different versions of the experiment. While some judges in the RP version of the experiment may object to realisations such as [nu:] for *new*, it would be pointless to present the same realisation in the GA version, where it is significantly less marked than [nju:]. This is why the latter has been selected as a potential “error” for the GA version of the experiment.

Because of such differences between versions, the effects of token similarity on native-speaker judges will be discussed in separate sections: one for the RP form and another one for the GA version.

Based on a representative selection of accent descriptions, Sections 4.2 and 4.4 describe, for all minor accent groups, which of the Dutch realisations represented by tokens 1 to 32 (excluding the distractor) are likely to be different from, or similar to, those produced by speakers of that accent group. In those cases where the majority of speakers in an accent group may be expected to pronounce a token differently from the Dutch English realisation offered in the experiment, the severity judgements for this particular token as provided by respondents in this group have been coded “DIFMAJ” (denoting a Dutch pronunciation that is **d**ifferent from the **m**ajority of speakers in this accent group). However, if the majority of judges in an accent group are likely to pronounce the token similarly to the Dutch English realisations produced by the actors, all severity judgements for this token given by participants in this group have been coded “SIMMAJ” (denoting a Dutch pronunciation **s**imilar to that of a **m**ajority of speakers in this accent group). Where such a claim can only safely be made about a minority of speakers, the code “SIMMIN” has been used (denoting a pronunciation **s**imilar to that of a **m**inority of speakers in this accent group). The “DIFMAJ”, “SIMMAJ” and “SIMMIN” codes are represented in Tables 4.1 and 4.2 (for the RP version of the experiment) and Table 4.3 (for the GA version of the experiment) – see 4.3 and 4.5.

4.2 Similarities to accents in the RP version of the experiment

4.2.1 BED

In the carrier sentence (see Table 2.26), the word-final voiced stop in *bed* is followed by a voiceless fricative. In this context, English /d/ (in all accents, standard or otherwise) is likely to be partly or even fully devoiced to [d̥] (Giegerich 1992: 222, Davenport & Hannahs 1998: 24), but will still be perceived as lenis (and hence in contrast with /t/) due to factors such as “vowel length, energy of articulation, lack of pre-glottalisation” (Collins & Mees 2003b: 52). While devoicing in English does not normally lead to a loss of phoneme contrast between /t/ and /d/, and is subject to variation depending on speaker, accent and context (Giegerich 1992: 223), Dutch devoicing is the result of the lack of any word-final phonemic contrast in obstruents.¹ Whereas some accents of English may be cited as having strong final devoicing – e.g. Standard Scottish

¹ There are studies which claim that Dutch has incomplete neutralisation in this context. For instance, Warner *et al.* (2004: 259) have conducted an acoustic investigation of final devoicing in Dutch, which, they argue, shows “reliable, if small effects of underlying voicing on [vowel] duration in the neutralization environment”.

English (Giegerich 1992: 223) and West Yorkshire (Davenport & Hannahs 1998: 24) – these sources do not specify whether this actually leads to a loss of fortis/lenis opposition.

For certain other accents, the indications are stronger. According to Wells (1982: 619), “devoicing of final voiced obstruents, leading to neutralization of the voicing opposition” may be heard in the English spoken by Afrikaans bilinguals, causing “*bed* and *bet*” to “become homophonous as [bet]”. In the English of Scots Gaelic bilinguals, the use of voiceless /b, d, g/ “may lead to their misidentification” as /p, t, k/, as evidenced by “eye-dialect” such as “*inteet* for *indeed*” found in literary sources from the mid-20th century (Wells 1982: 413). In her description of Highland and Island English, Shuken (1984: 156) also notes that “the voiced/voiceless distinction of stops does not seem to be ... consistently maintained”. In other varieties of English, loss of fortis/lenis contrast for /t ~ d/ in word-final position may also occur in very specific contexts or as the result of assimilation, but these variable, context-specific phenomena are found in a limited number of speakers and are thus less likely to affect most judges’ perception of word-final Dutch substitution of /d/ by /t/. A well-known example is Yorkshire Assimilation (Wells 1982: 366). Penhallurick (2004: 109) also refers to the occasional “unvoicing” of final /d/ in Welsh English.

4.2.2 BAT

In a number of accents, the TRAP vowel is very commonly realised as [ɛ], which is strikingly similar to the realisation of English /æ/ by native speakers of Standard Dutch.² These accents include London & the Southeast, Australian, New Zealand and South African English (Wells 1982: 305, 598, 607, 613). According to Lass (1990: 276), this realisation is characteristic of “vernacular” varieties of South African English rather than of “posher speakers”, whose realisation is “somewhat more [æ]-like”. Some speakers of other accents may also realise the TRAP vowel as [ɛ] – but normally only in the environment of adjacent velars, as in some middle-class varieties of Scottish English and in some varieties of Northern Irish English (notably Belfast), or as a lexical-incident phenomenon in Southern Irish (Wells 1982: 403, 442, 423; see also Hickey 2004b: 73). Even though some of these speakers may pronounce *cattle* as [ˈkɛtɫ] or *carry* as [ˈkɛrɪ], no speaker of any variety of Scottish or Irish English would normally realise *bat* as [bet].

For all other native speakers of English, TRAP and DRESS never appear to occupy the same phonological space in similar phonological contexts. Since this sample contains no example of the DRESS vowel, judges will be unable to normalise for this contrast on the basis of this utterance alone. (They may, of course, take into account other samples of the RP actor’s speech, which most of

² As in Wells (1982), the symbol e has been used for the vowel in RP DRESS, and ɛ for the DRESS vowel in GA. The symbol ɛ has been employed for any front open-mid realisations close to CV3. See also 4.4.2.

the respondents will have previously encountered in the course of participating in the experiment. In the absence of any other instantiations of raised front vowels, judges are unlikely to normalise for this.)³

4.2.3 VAN

Substitution of /v/ by /f/ does not occur in any of the relevant accents, with one – admittedly questionable – exception. For Gaelic-influenced English, Wells (1982: 413) gives an example taken from a mid-20th century novel which would suggest initial /f ~ v/ confusion: “ferry good”. According to Shuken’s (1984: 156–158) more recent description of Highland and Island English, the “voiced/voiceless distinction of stops and fricatives does not seem to be as consistently maintained” in Hebridean English “as it is in many other accents of English”, but this appears to be less frequent and less perceptually important for fricatives in word-initial position. Nonetheless, the only judge in this survey to label himself a speaker of “Scottish Gaelic English” states that *fan* for *van* is a “Gaelic pronunciation” (Subject 852). As a result, /f ~ v/ confusion has been labelled as occurring in a minority of speakers of Scots Gaelic.

A related phenomenon, /f/ voicing (arguably the opposite of /f/ substitution), which was traditionally found in the West Country of England, is now so “archaic” as to be unlikely to influence any listeners (Wakelin 1984: 75). It was also formerly found in a few peripheral areas of South Wales (Penhallurick 2004: 109), as a result of settlement from the West Country.

4.2.4 WINE

Dutch /v/ is often used as a substitution for English /w/ and may be perceived by native speakers as /v/ (Collins & Mees 2003b: 174–175). Confusion of /w ~ v/ is not attested in any of the relevant accents, and has therefore been excluded from further consideration. For instance, according to Wells (1982: 332–333), the “supposed interchange of [v] and [w]” as found in “literary representations of Cockney speech” is now “utterly obsolete”; see also Ellis 1889: 132, 229, Ihalainen 1994: 227, Trudgill 2004: 174–175.) Interestingly, Trudgill (2005b: 214) states that accents “such as those of Tristan da Cunha and the Bahamas ... lack a contrast between /w/ and /v/”. Similarly, Wells (1982: 568) notes that “Bahamians, Bermudans, and Vincentians are among those for whom the use of [w] for standard [v], or a bilabial fricative [β] for both, has been reported”. However, no respondents identified themselves as hailing from these islands – nor did any participants describe themselves as speakers of Indian English (which is another variety in which the contrast is not commonly made, see Wells 1982: 629).

³ On speaker normalisation, see Johnson (2005).

4.2.5 THIN, AUTHOR, BOTH, THAT, WEATHER, BREATHE

Accents which have TH-fronting (replacement of dental by labio-dental fricatives /f v/) as opposed to TH-stopping (replacement of dental fricatives by dental plosives [t d] or affricates [tθ dð]) have been categorised as being different from Dutch English. Even though speakers of those accents may well be more lenient towards other non-standard native or non-native realisations of /θ/ and /ð/, it would be impractical to test this in the present survey without expanding the experimental set-up excessively to include realisations that are rarely found in Dutch learners.

Some speakers of Southern Irish English have realisations of /θ/ and /ð/ which are similar to the Dutch English substitutions of alveolar [t] and [d], but these are associated with “traditional rural ... varieties of Irish English, especially in the east, south and southwest of the country” (Hickey 2004a: 81) and are “highly stigmatised” (Hickey 2004b: 92; see also Hickey 2004a: 59). Most Irish people would use dental [t] or [d] instead, or would use a range of other possibilities (see Wells 1982: 429, Hickey 2004b: 75). This means that only a minority have similarities to Dutch English in this respect. The latter is also true of a small group of speakers in Northern England, notably “working-class Catholics” from Liverpool (Wells 1982: 371).

There appears to be little clarity as to the situation in the Scottish Highlands and Islands. As Wells (1982: 413) points out, “Gaelic does lack ... phonemes corresponding to English /θ/, /ð/, and /w/”, but nevertheless substitutions “such as [sɪŋk] for *think* are not found in the post-Gaelic Highlands”. However, in her article on Hebridean English, Shuken (1985: 149) states that /θ/ and /ð/ are “occasionally ... pronounced as dental stops, although this is rare”. In any event, the only judge in this survey to label himself a speaker of “Scottish Gaelic English” states that “[e]veryone in my part of Scotland has trouble with ‘th’”, presumably referring both to /θ/ and /ð/. In fact, this respondent refers to TH-stopping as a “Gaelic pronunciation” (Subject 852). This is why TH-stopping has here been categorised as occurring in (at least) a minority of speakers of Scots Gaelic. It is also found in Shetland and to a certain extent in Orkney (McClure 1994: 67, Melchers 2004: 42, Van Leyden 2004: 20, Wells 1982: 399), but none of the participants identified themselves as hailing from these islands. “Medial /ð/”, according to McClure (1994: 66), “is replaced by /d/” in the north-east of Scotland, and “occurs occasionally in Scots in Glasgow” (Stuart-Smith 2004: 62). In other words, /ð/ substitution has been attested in a minority of speakers of Scottish English, and is presumably restricted to this environment.

In one specific environment, Londoners may also realise /ð/ as [d]: “[w]ord-initially after a word ending in a consonant” (Wells 1982: 329). This is the environment provided in the carrier sentence for THAT: “We were supposed to be meeting that man at two o’clock”. According to Wakelin (1984: 79), this is found in words such as “*that, the, there and these* ... in popular London dialect and the south-east”. There is no evidence to suggest that this is found in a majority of Londoners.

In addition, it is a well-known fact that TH-stopping is a salient characteristic of virtually all varieties of West Indian English (see Aceto 2004: 486, Devonish & Harry 2004: 475, Schneider 2004a: 1085), but no speakers of these appear to have taken part in the experiment.

4.2.6 OFF

Even if substitution of /v/ by /f/ before a consonant is rare, an example may be found in the standard RP pronunciation of *have to* as /hæftə/ (Collins & Mees 2003b: 210). Substitution of /v/ by /f/ before a vowel, however, as in sentence 10 (“Many of our students come from English-speaking countries”), is not attested in any of the relevant accents. This is probably also true of the English spoken in the Scottish Highlands and Islands, where final devoicing of fricatives is particularly found “before pause and before voiceless consonants, but even sometimes before voiced consonants” (Shuken 1984: 158). This may be taken to mean that it is far less frequent, or possibly even absent, before vowels.

4.2.7 RED

According to Aitken (1984: 102), uvular-*r* is found in Scotland “in a sizeable minority of speakers, not apparently local to any one area”, and Wells (1982: 411) terms it a “surprisingly common ... personal idiosyncrasy” which “can hardly be regarded as a local-accent feature”. In Wales, “it is to be heard in parts of both Gwynedd and Dyfed as a social/geographical characteristic, not just as a personal idiosyncrasy” (Wells 1982: 390) – although Penhallurick (2004: 110) states that such “realizations ... are confined to the north, where they are rare and possibly usually idiolectal”.

Similarly, uvular articulations are heard in some Northeastern English, notably as a salient feature of the traditional speech of Tyneside (Wells 1982: 368). Beal (2004: 129) points out that this pronunciation, “known as the Northumbrian burr, ... has been a source of pride to Northumbrians, many of whom today will perform the burr as a party-trick even though they would not use it in everyday speech”. For Southern Ireland, Wells (1982: 432) reports that there is “some suggestion that /r/ may have velar or uvular variants in County Louth and County Tipperary/County Limerick”. According to Hickey (2004a: 79), “uvular [ʁ] extends across north Leinster under the border with the north and is found in Cavan”.

Even though in South Africa pre-vocalic alveolar trills are common in Afrikaans bilinguals, uvular-*r* is much less widespread: it is popularly supposed to be characteristic of speakers from Malmesbury (Wells 1982: 617). It is therefore likely that a uvular trill such as the present speaker’s will only be found in a minority of Afrikaans speakers.

4.2.8 ICE

Although realisations of /aɪ/ differ widely across the English-speaking world, lengthening of the first element of this diphthong before a fortis consonant (tautomorphic or otherwise) is not attested in any of the accents. (See, for

instance, Schneider 2004a: 1118 for a discussion of different varieties of the PRICE vowel.)

4.2.9 TIE

Aspiration of stressed syllable-initial fortis plosives (or in the case of /t/, possibly more frequently affrication) is normal in Standard Southern British English, the only doubtful case being the very rare “upper-crust RP”, which, according to Wells (1982: 282), often has “surprisingly little aspiration”. A minority of speakers of Northern British English have non-aspiration in this environment, notably from “the Pennine valleys north of Manchester” (Wells 1982: 370). The same is true for “many speakers” of Scottish English, with the notable exception of the Highlands and Islands (Wells 1982: 409). Non-aspiration is also a feature of some South African English, but it is “reported to be receding” even though it remains “unstigmatized” (Wells 1982: 618). (See also Bowerman 2004: 939 on non-aspiration in White South African English, and Van Rooy 2004: 949–950 on aspiration in Black South African English.)

Wells does not discuss non-aspiration with reference to Irish English, Australian English or New Zealand English, presumably because it is not a salient characteristic of these varieties. In fact, Horvath (2004: 635) states that in Australian English, “plosive [t] has the usual allophonic distribution for the aspirated and unaspirated variants”, while Bauer & Warren (2004: 593) mention that in New Zealand English, “alveolar [t] is affricated initially in stressed syllables” and “bilabial [p] can be heard aspirated in all positions”. According to Gimson & Cruttenden (1994: 151), however, “Irish English and Welsh English” have even “more aspiration than RP” in initial /p, t, k/. Hickey (2004b: 93), for instance, refers to “Fashionable Dublin English speakers” who “may have slight affrication of syllable-initial /t-/, as in *two* [tsu:]”. In his chapter on Welsh English, Wells (1982: 388) also notes that these sounds “are strongly aspirated in most positions”. (See also Penhallurick 2004: 108–109.) This would suggest that non-aspiration is unlikely to occur in either Irish English or Welsh English.

4.2.10 DEAD

According to Wells (1982: 326–327), speakers of Cockney may sometimes use [ʔ] for /d/ before a word or syllable boundary followed by a consonant, as in [eʔwəʔ 'kʌɪm] *Edward came*. This is the same environment as *dead with* in the present survey. Wells states that this realisation is less “common” and possibly “lexically-restricted” (Wells 1982: 326–327).

4.2.11 FILM

Schwa epenthesis is found in a number of accents. According to Wells (1982: 435), it is found in the “popular speech” of Irish English. It is also cited by Todd (1992: 68) as one of the features of Anglo-Irish that are not exclusively found in the South. Hickey (2004a: 83) also confirms that “epenthesis” as “in the Irish English pronunciation of the word *film* as [fɪləm]” is to be found “across the entire country”. According to McArthur (1992b: 336), it is also to be heard in

Scottish English: “[w]orking-class Edinburgh speech shares features with Glasgow and other Central Scots dialects” such as “the epenthetic vowels in ‘girrul’ for *girl* and ‘fillum’ for *film*”. (See also Burchfield 1994b: 12.) Shuken (1984: 160) mentions this as a characteristic of Scots Gaelic, where “[s]ome speakers have an epenthetic vowel between /l/ and a following nasal (e.g. in ... film [fɪlɪm], [fɪləm])”. She also cites the example of “an epenthetic central vowel between /r/ and /m/ in words like *arm*”, which is also found “in some Lowland varieties of Scottish English” (Shuken 1984: 160).

A number of judges from Northeastern England stated that schwa epenthesis is also a well-known feature of the local accents of that region. Collins & Mees (2003b: 171) mention Lancashire as a case in point, while Beal (2004: 130) provides examples from Tyneside and Northumberland. According to Branford (1994: 485–486), it is also “common” in “broader South African English”, either as a result of influence from Afrikaans or because it already existed in the “British rural dialects” of the English-speaking settlers. Other places where this pronunciation may be found include Sydney (Burchfield 1994b: 12) and New York (Branford 1994: 486).

4.2.12 CAR

Some judges, especially those speaking non-rhotic varieties of English, will object to the presence of post-vocalic /r/, whereas others, notably speakers of rhotic varieties, could conceivably have objections to this particular realisation of /r/ (or to its length).

Most of England is non-rhotic (including Greater London and the Midlands), but in the “west and north-west” there are areas with rhotic or variably rhotic speakers (Wells 1982: 220). In the West Country, where “[f]ull rhoticity ... extends well up the social scale” (Wells 1982: 341), there are considerable numbers of rhotic speakers, but if we accept Wells’s claim that “middle-class accents and, increasingly, working-class accents of the traditionally rhotic areas of the west ... of England now tend to exhibit no more than variable rhoticity”, these must be a minority, if not within their own areas, then certainly within Southern England as a whole (Wells 1982: 220). Rhotic speakers are also to be found in Lancashire and other parts of Northern England, but these are far from being a majority (Wells 1982: 367–368). The latter is also true of Wales (Wells 1982: 378).

Since prestige accents in England and Wales are non-rhotic, there appears to be considerable stigmatisation of rhoticity in these countries, even in rhotic areas. “Rhotic speech”, according to Collins & Mees (2003b: 180–181), “is frequently employed on the British stage for comic effect, and is thought of as being characteristic of ‘rustic’ or ‘peasant’ dialect ...”. Collins & Mees state this is particularly true of speech “with a strong post-vocalic retroflex type /r/ similar to that used by many Dutch learners” (2003b: 180). According to Altendorf & Watt, this “feature is perceived as particularly pleasing by many speakers from outside the area, but is at the same time one of the major stereotypes responsible for the impression of rusticity also often associated with Southwestern accents”

(2004: 197). This may affect English and Welsh judges' reactions to this feature of Dutch English.

Apart from "a shrinking area of rural Southland", New Zealand is non-rhotic (Bauer 1994: 411; see also Bauer & Warren 2004: 594), as are "all native-speaking accents of South Africa and Australia" (Wells 1982: 220). "There is", according to Bowerman (2004: 940), "some evidence of post-vocalic /r/ in some Broad Cape varieties", but none of the respondents have identified themselves as speaking this variety.

As in the US and Canada, rhoticity is the norm in both Scotland and Ireland (Wells 1982: 407, 432) – an interesting exception being working-class Dublin speech (Hickey 1999: 272, 2004b: 92). The examples provided in Shuken (1984: 164) demonstrate that the English spoken in the Scottish Highlands and Islands is rhotic, too. This would suggest that Scottish and Irish judges are very unlikely to object to rhotic distribution patterns.

In the experiment, the RP actor's realisation of post-vocalic /r/ is a bunched palatal approximant, the auditory effect of which appears to be hardly or no different from that of the "retroflex type /r/" heard both in Dutch English (Collins & Mees 2003b: 180) and in some other rhotic accents, including the West Country (Wells 1982: 342) and the recessive rhotic areas of Dyfed, Gwent and Powys (1994: 131). Needless to say, any such realisation is untypical of the majority of speakers in Southern England and Wales, who do not pronounce post-vocalic /r/ at all – and whose attitude to any type of retroflex /r/ may well be biased.

There is no evidence to suggest that either bunched palatal approximant or retroflex /r/ are commonly found in rhotic areas in England and Wales other than the areas mentioned above. According to Wells (1982: 368), a "post-alveolar approximant" is "usual" in Northern England – alongside with "alveolar tap", whereas an "alveolar roll ... or tap" is found in the "second-language English of those who have Welsh as their first language" (Wells 1982: 378–379). Nor are such realisations likely to occur in the rhotic accents of southernmost New Zealand. Little is known about post-vocalic /r/ in the Southland, but it is reported to be no different from /r/ in other parts of the country (Wells 1982: 606), and therefore hardly likely to be a retroflex approximant (although Gordon & Maclagan 2004: 606 refer to research by Bartlett which "indicates that the realisation of postvocalic /r/" in Southland English "is approximal rather than rolled or flapped"). In any event, as post-vocalic /r/ is only found in a minority of speakers, its realisation (whatever it is) must be regarded as untypical of the majority of New Zealanders, whose speech is non-rhotic. This is also true of Australians. Even though Trudgill & Hannah (2002: 18) describe Australian /r/ as "often more strongly retroflexed" than in the English spoken in England, any such realisation of post-vocalic /r/ must be considered uncharacteristic of Australian English.

Neither bunched palatal approximant nor retroflex approximant are typical of the majority of Scottish speakers of English. Wells (1982: 411) associates post-vocalic /r/ in Scottish English with "a post-alveolar or retroflex

fricative or approximant”, which would suggest that a retroflex approximant realisation is just one of different options. The bunched palatal approximant was characterised as “non-Scottish” by one Scottish judge, who stated that it sounded “like southwest England” (Subject 852). Neither are such realisations typical of English in the Scottish Highlands and Islands, where the most common word-final realisation of /r/ is what Shuken (1984: 160) describes as “a fricative, or an affricated tap (a tap followed by a fricative)”.

In two other rhotic areas, however, post-vocalic /r/ is actually very commonly realised as a retroflex approximant [ɻ]: Northern Ireland and the Irish Republic (Wells 1982: 446, 432).⁴ According to Hickey (2004b: 87), retroflex [ɻ] is, for instance, to be found in “current fashionable Dublin English” and is “widespread throughout Ireland among younger female speakers” (see also Hickey 2004a: 78). This would suggest that Irish judges are much less likely to object to this realisation than any other group of respondents.

4.2.13 HOT_TEA

This error is an example of degemination across word boundaries. Whilst some accents have gemination with compensatory lengthening (Wells 1982: 388 provides the example of “/-tt-/ = [t:]” in RP), and other accents may have glottal replacement, i.e. /-tt/ = [ʔt], the only accent reported to have any possible “loss of opposition between single and double consonants” is Welsh English, implying that both *meeting* and *meat-tin* may be pronounced as [mi:t:m] (Wells 1982: 388). This, however, is the opposite effect of the degemination found in Dutch English, as in [ʰoti:], and therefore unlikely to affect the assessment of any judges.

4.2.14 NEW

Yod-dropping is found in speakers from Southern England and the Midlands, notably in what Hughes & Trudgill (1987: 36) describe as “a large area of eastern England”. Wells (1982: 330, 338) states that it is “commonly heard in working-class London speech” in “the environment /n_/” and “widespread” in East Anglia in “virtually all environments except word-initial”. Its occurrence in words such as *new* is also attested for the West Midlands, especially in “teenage speech” (Mathisen 1999: 111; see also Clark 2004: 158). Judges frequently referred to it as “American”. In addition, [nju] may be realised as [nru] in most of South Wales (see also Harris 1994: 87, Wells 1982: 385). This is an example of the GOOSE/JUICE split which Collins & Mees describe as being typical of all Welsh English except for the Cardiff/Newport area (2003b: 303). In any event, the difference between the two realisations is only “a minor one phonetically”

⁴ As far as American English is concerned, it should be noted that according to Higgs’s (1980: 116) re-evaluation of a study by Delattre, the retroflex approximant is rare in post-vocalic position. In the study discussed by Higgs, “less than 4%” of post-vocalic realisations are “apical retroflex or alveolar, and a staggering 86% are dorsal or labial advanced velar”.

(Wells 1982: 385). It is therefore unlikely to bias any Welsh listeners differently from those who say [nju:]. Schneider (2004a: 1124) also includes Irish English among yod-dropping varieties of English, but no other sources which conclusively confirm this have been found.⁵ Variable yod-dropping “following /n/” has also been attested for New Zealand (Bauer 1994: 388), but is mainly associated with “high frequency collocations” such as “New Zealand” and “Air New Zealand” (Bauer & Warren 2004: 597).

4.2.15 IMAGIN

There is not enough reliable data to suggest that [ɪmædʒmeɪtɪv] is an acceptable minority pronunciation in Britain. This also applies to [ɪmædʒə'neɪtɪv] in North America. Although neither the *Longman pronunciation dictionary* (Wells 2000) nor the *English pronouncing dictionary* (Jones 2003) record [ɪmædʒmeɪtɪv] for RP, two non-British judges of the RP version of the experiment suggested it may be a British pronunciation. Whilst virtually all respondents identified it as an error and many stated that it made the speaker sound non-native or foreign, one participant said it made the speaker sound “stupid” (Subject 811) and two judges referred to it as a “class indicator” (Subjects 313 and 852). One of these also pointed out that [ɪmædʒmeɪtɪv] “can be heard throughout the world of new [E]nglishes” (Subject 313). This contention may be difficult to prove, and it should be noted that the different stress patterns found in such varieties as Indian English are regularly interpreted as errors by speakers of other varieties (Wells 1982: 630, Trudgill & Hannah 2002: 130). In any event, the only variety for which [ɪmædʒə'neɪtɪv] is recorded is as an alternative pronunciation for GA. One of the Australian judges of the British version appears to be alluding to this when stating that it “sounds odd” to the “native speaker” to have the “fourth syllable stressed (except for perhaps some regional varieties of American English” (Subject 999). In the GA version of the experiment, however, the GA speaker actually says [ɪmædʒə'neɪtɪv], with the stress on the penultimate, which none of the respondents recognise as an acceptable regional variation.

4.2.16 PERFECT

There is no evidence to suggest that the adjective *perfect* is normally pronounced with final stress in any variety of native-speaker English. Even if word stress in varieties such as Indian English is strikingly different from the patterns found in RP or GA – making final stress more likely in this instance – such realisations are commonly interpreted as errors by other native speakers (Wells 1982: 630, Trudgill & Hannah 2002: 130). At any rate, none of the judges described this pronunciation as being characteristic of another variety of English

⁵ For instance, Hickey (2004a: 83) states that “there is widespread deletion of yod in a position following an alveolar sonorant in stressed position, e.g. news [nu:z]” but the recordings of different varieties of Irish English provided on the accompanying DVD do not suggest that this is found in most accents.

– the sole exception being a self-styled speaker of BBC English who thought that this was an “American pron[u]nciation” (Subject 929).

4.2.17 TO_WALES, THAT_THA, WOULD_ON

These examples illustrate the overuse of Strong Forms (SFs) in words such as *to*, *that* and *would*, in conjunction with incorrect stressing. According to Collins & Mees (2003b: 20), “the excessive use of SFs is one of the main sources of error for Dutch-speaking students of English”. In most utterances, native speakers of English normally use the corresponding Weak Forms (WFs) instead (Gimson & Cruttenden 1994: 230). In fact, it would be difficult to find any evidence for Jenkins’s (2004: 146–147) claim that “speakers of certain L1 varieties (for example, Scottish and South African English)” “regularly fail to produce weak forms”. It is true that in “popular Scottish English there are certain weak forms sharply distinct phonetically from those found outside Scotland”, such as “[tə] or [tɪ] ... as against the [tu] ~ [tʊ] ~ [tə] of English accents” (Wells 1982: 414), but even if these realisations are different from RP, there is no evidence to suggest that Scottish or South African speakers fail to use them in appropriate contexts.

It may well be possible to find examples of weak form avoidance in certain creoles and pidgins, but since these varieties of English are not represented by any of the judges, such considerations must fall outside the scope of this experiment. WF avoidance is also reported to be common in Indian English, where, according to Wells (1982: 627), weak forms are not “regularly used”: “to is always [tʊ]”.

In the case of [ðæt ~ dæt], three judges stated that the pronunciation of the sentence reminded them of Yorkshire or Northern England. Since none of the judges appeared to hail from this part of Britain, it would be difficult to accept their claims at face value. In his discussion of the STRUT words, Wells (1982: 352) also notes a “lack of distinction” in “northern Near-RP”, for instance “between the strong and weak forms of *but*, *does*, *must*, *us*”, but this does not necessarily apply to *that* (the strong form of which has a TRAP rather than a STRUT vowel). At the same time, Wells states that the word *that* “is among those with no distinct weak forms in Tyneside speech” – which would make this a minority realisation in Northern England (Wells 1982: 376).

4.2.18 SECONDAR

Some judges described the RP actor’s pronunciation of *secondary* as American – not surprisingly since [ˈsekəndəri] is indeed the standard pronunciation in General American. In addition, the *Longman pronunciation dictionary* (Wells 2000) lists [ˈsekəndəri] as a “non-RP” variant pronunciation in British English, without linking it to any specific accent. Nor did RP judges who recognised this variant pronunciation attempt to localise it. One Australian judge described it as a “perfectly acceptable pronunciation ... although amongst British, Irish and Australian speakers the ‘a’ is dropped – particularly speakers over 40 years” (Subject 999). Many judges from these countries did indeed object to this pronunciation, preferring the weakened or elided penultimate vowel commonly

found in RP (Wells 1982: 231). Wells describes this pronunciation as an RP “innovation”, which “has hence also, to a varying extent, affected other British accents and those of the southern hemisphere”. Unfortunately, Wells does not further expand on this “extent” – except to say that “near-RP often preserves a strong vowel” in words such as *necessary* – and therefore presumably also in similar words (Wells 1982: 231). If this is indeed an RP-led innovation which has not yet been completed, as Wells seems to be suggesting, there will be at least a minority of speakers, in all the accents affected, who will preserve a strong vowel in *secondary*. To support this view, it may be argued that weakening is a common natural phonological process, found in many varieties of English, and one which might be rejected by upwardly mobile and linguistically insecure speakers who are possibly unduly influenced by orthography.

4.2.19 TELL

Even though final *l* in /*tel*/ would be realised as [ɫ] in most varieties of English, a Dutch realisation of dark [ɫ] may attract attention because, according to Collins & Mees, it has “pharyngealisation rather than velarisation with a noticeable retraction of the tongue-root” (2003b: 170). Little is known about the effects of such an overdark [ɫ] on native speakers of different varieties of English. Interestingly, Wells (1982: 603, 609) suggests that [ɫ] in Australian and New Zealand English may in fact be pharyngealised. In addition, Bauer & Warren (2004: 595) cite what would appear to be a pharyngealised “[ɫ̠]” (presumably [ɫ̠]) as a possible allophone of /*l*/ in New Zealand English. If these are common realisations, speakers of these varieties are much less likely to object to overdark Dutch [ɫ] than speakers whose [ɫ] is velarised, as in RP (Collins & Mees 2003b: 169) or most Scottish English (Wells 1982: 411). For speakers of the latter varieties, social stigma may even be attached to pharyngealised realisations of [ɫ], as is true of the [ɫ] heard in the “popular speech” of the “Glasgow area” (Wells 1982: 411; see also Stuart-Smith 2004: 63). Another area where pharyngealised [ɫ] may be heard is North Wales (Wells 1982: 390, Penhallurick 2004: 110). Clear [l], even word-finally, is common in Ireland – although much less so with “young female speakers” (Hickey 2004b: 87). In fact, Hickey (2004a: 77) confirms that “more recent supraregional varieties of English in the south of Ireland all show a clearly velarised /*l*/ in syllable-final position”. A clear realisation of word-final [l] is also widespread in the Scottish Highlands and Islands, most of South Wales, South Africa and parts of Northern England. Speakers of such varieties are as likely to object to pharyngealised [ɫ] as those whose varieties have velar [ɫ] – if not more.

Another feature of dark [ɫ] in Dutch is, according to Collins & Mees, its vowel-like articulation (2003b: 171). This may be similar to the L-vocalisation heard, according to Wells, in “Near-RP” and “London, where it is overtly stigmatized” (Wells 1982: 295, 314). More recently, Altendorf & Watt (2004: 196) have reported that L-vocalisation “is spreading regionally, so far mostly within the Southeast, and socially to higher social classes. In London, Kent and Essex ... it is already very frequent, almost categorical, in the accents of younger

middle-class speakers". Trudgill (2004: 175) also states that it is "increasingly common" in the southern part of East Anglia. In addition, it has been attested for Glasgow, "especially for working class adolescents" (Stuart-Smith 2004: 63). Horvath (2004: 641) considers that L-vocalisation "in London English does appear to be comparable [to Australian English]". It has also been reported for New Zealand (Wells 1982: 609, Bauer & Warren 2004: 595, Gordon & Maclagan 2004: 611).

In any event, it is difficult to equate the realisation of Dutch [ɫ] with anything found in native accents, although one listener was reminded of "a Lancashire 'L'" (Subject 642). This may be because "dark [ɫ] may occur in all positions" "in large parts of the North of England (e.g. Manchester)" (Gimson & Cruttenden 1994: 184). In view of this difficulty, it would seem prudent to categorise any variety that has either pharyngealised or vocalised [ɫ] as being only similar to Dutch English for a minority of speakers.

4.2.20 COLOUR

This is a common spelling pronunciation found in Dutch learners. A number of non-Northern judges also associated this pronunciation with the north of England. Although realisations such as [kʊlə] or [kɪlə] would be more characteristic of Northern English (see Wells 1982: 356–362), there are indeed areas in the North and the Midlands where *one* is pronounced /wɒn/ and where "once, among, none and nothing may also be encountered with /ɒ/" (Wells 1982: 362). These would appear to be cases of lexical incidence, but examples such as "money, slush, other, mother", which are "occasionally" found in the Sheffield dialect, "particularly with females of middle age" (Stoddart *et al.* 1999: 74), suggest wider membership for this lexical set (at least for those speakers). Interestingly, the "most common variant" of /ʌ/ is reported to be [ɒ] "for all generations of speakers" in Sandwell, a borough in the West Midlands close to Birmingham (Mathisen 1999: 108). Mathisen claims that it is to be found "especially in monosyllabic words where most Northern varieties have [ʊ]" and "occurs very frequently with the elderly, in all phonetic contexts", whereas for "younger speakers, it is more frequent before /l/ and /ŋ/" – as in *colour* (Mathisen 1999: 108).

According to Wells (1982: 132, 422), /ʌ/ may be pronounced as "[ɔ]" in "Irish and West Indian accents" – other Irish realisations are "of the type [ɪ, ə]" or are "indistinguishable from conservative RP [ʌ]". The realisation [ɔ] is also heard in the "Anglo-Irish" area of Northern Ireland (Wells 1982: 442) and in Derry (see McCafferty 1999: 246). Such variation makes it more difficult to predict if any Irish judges, whether from the North or the South, perceived the RP actor's [kʊlə] as similar to Irish realisations. It would be safe to assume that this will be the case for at least a minority of speakers.

4.2.21 STOOD

The contrast between /ʊ/ in FOOT and /u:/ in GOOSE is not commonly made by Dutch learners; whilst beginning Dutch learners of English may substitute /u/ for

both (often realised as a somewhat fronted [ʊ]), more advanced learners tend to use either /ʌ/ or an “extended glide of an /ʌy/ type” similar to the RP actor’s realisation (Collins & Mees 2003b: 97). Interestingly, a similar “FOOT-GOOSE merger is characteristic of all Scottish accents of all regional and social types; but of no others, except only those of Ulster and northernmost Northumberland” (Wells 1982: 402). Nor do the Gaelic speakers studied by Shuken make the contrast (1984: 162). In addition, the most common Scottish realisation of the vowel in either FOOT or GOOSE is in fact [ʌ] (Wells 1982: 402). The same would also appear to be true of Ulster (Wells 1982: 441, Hickey 2004b: 91) and some of the Gaelic speakers studied by Shuken (1984: 163). All this would suggest that the vast majority of Scottish and Northern Irish speakers (as well as perhaps a tiny minority of speakers from Northern England) are unlikely to object to the RP actor’s realisation of /ʊ/.

In some accents, notably “in Ireland and parts of the north of England” FOOT words spelt *-ook* may be pronounced with a GOOSE vowel (Wells 1982: 133). This can possibly be the reason why two judges thought this RP actor’s realisation of /ʊ/ sounded like a “Yorkshire” accent (Subjects 358 and 956). There is no evidence to suggest, however, that any of those accents have a GOOSE vowel in a word like *stood* (which does not have the requisite *-ook* spelling).

4.2.22 INT1, INT2, INT3

It is difficult to instruct bilingual actors to use convincing Dutch intonation patterns in a segmentally correct English sentence. This is why the actors were asked instead to use authentic English intonation patterns, the pitch contours of which were subsequently removed using the speech manipulation program PRAAT (Boersma & Weenink 2002) and replaced by the intonation contours produced by a Dutch speaker reading out phrases with comparable length and nucleus location (see 2.4.3). This resulted in segmentally correct English sentences with authentic Dutch intonation contours, which were presented to the judges. Since it would appear highly unlikely that any native variety of English has exactly the same intonation patterns as Dutch, it would seem improbable that any listeners would be affected by this and that they would judge Dutch intonation patterns more leniently as a result.

4.3 Accent similarity codes for the RP version of the experiment

On the basis of the discussion in 4.2, similarity codes have been assigned to all severity judgements in any one minor accent group taking the RP version of the experiment. This is shown in Tables 4.1 and 4.2. The “SIMMIN” and “SIMMAJ” codes are in bold.

Table 4.1. Accent similarity codes for all minor accent groups taking the RP version of the experiment (by token; British groups only).

	<i>GB/RP</i>	<i>GB/LO</i>	<i>GB/SO</i>	<i>GB/MI</i>
<i>BED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BAT</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ
<i>VAN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WINE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>AUTHOR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BOTH</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>OFF</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ
<i>WEATHER</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BREATHE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>RED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>ICE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TIE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>DEAD</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ
<i>FILM</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>CAR</i>	DIFMAJ	DIFMAJ	SIMMIN	DIFMAJ
<i>HOT TEA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>NEW</i>	DIFMAJ	SIMMIN	SIMMIN	SIMMIN
<i>IMAGIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>PERFECT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TO WALES</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT THA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>SECONDAR</i>	DIFMAJ	SIMMIN	SIMMIN	SIMMIN
<i>WOULD ON</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TELL</i>	DIFMAJ	SIMMIN	SIMMIN	DIFMAJ
<i>COLOUR</i>	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN
<i>STOOD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT1</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT2</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT3</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

Table 4.1 (continued).

	<i>GB/NO</i>	<i>GB/WA</i>	<i>GB/SC</i>	<i>GB/SG</i>
<i>BED</i>	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN
<i>BAT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>VAN</i>	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN
<i>WINE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THIN</i>	SIMMIN	DIFMAJ	DIFMAJ	SIMMIN
<i>AUTHOR</i>	SIMMIN	DIFMAJ	DIFMAJ	SIMMIN
<i>BOTH</i>	SIMMIN	DIFMAJ	DIFMAJ	SIMMIN
<i>OFF</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT</i>	SIMMIN	DIFMAJ	DIFMAJ	SIMMIN
<i>WEATHER</i>	SIMMIN	DIFMAJ	SIMMIN	SIMMIN
<i>BREATHE</i>	SIMMIN	DIFMAJ	DIFMAJ	SIMMIN
<i>RED</i>	SIMMIN	SIMMIN	SIMMIN	DIFMAJ
<i>ICE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TIE</i>	SIMMIN	DIFMAJ	SIMMAJ	DIFMAJ
<i>DEAD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>FILM</i>	SIMMIN	DIFMAJ	SIMMIN	SIMMIN
<i>CAR</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>HOT TEA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>NEW</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>IMAGIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>PERFECT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TO WALES</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT THA</i>	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>SECONDAR</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>WOULD ON</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TELL</i>	SIMMIN	SIMMIN	SIMMIN	DIFMAJ
<i>COLOUR</i>	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>STOOD</i>	SIMMIN	DIFMAJ	SIMMAJ	SIMMAJ
<i>INT1</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT2</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT3</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

Table 4.2. Accent similarity codes for all minor accent groups taking the RP version of the experiment (by token; non-British groups only).

	<i>IRL/N</i>	<i>IRL/S</i>	<i>SA</i>	<i>NZ</i>	<i>AU</i>
<i>BED</i>	DIFMAJ	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ
<i>BAT</i>	DIFMAJ	DIFMAJ	SIMMAJ	SIMMAJ	SIMMAJ
<i>VAN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WINE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THIN</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>AUTHOR</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>BOTH</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>OFF</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>WEATHER</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>BREATHE</i>	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>RED</i>	DIFMAJ	SIMMIN	SIMMIN	DIFMAJ	DIFMAJ
<i>ICE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TIE</i>	DIFMAJ	DIFMAJ	SIMMIN	DIFMAJ	DIFMAJ
<i>DEAD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>FILM</i>	SIMMIN	SIMMIN	SIMMIN	DIFMAJ	DIFMAJ
<i>CAR</i>	SIMMAJ	SIMMAJ	DIFMAJ	SIMMIN	DIFMAJ
<i>HOT TEA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>NEW</i>	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN	DIFMAJ
<i>IMAGIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>PERFECT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TO WALES</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT THA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>SECONDAR</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>WOULD ON</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TELL</i>	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN	SIMMIN
<i>COLOUR</i>	SIMMIN	SIMMIN	DIFMAJ	DIFMAJ	DIFMAJ
<i>STOOD</i>	SIMMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT1</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT2</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT3</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

4.4 Similarities to accents in the GA version of the experiment

4.4.1 BED

Loss of fortis/lenis contrast for /t~d/ in word-final position has not been attested unambiguously in any native-speaker varieties of North American English. It is true that in African American Vernacular English or Black English, “[s]ingle final /b, d, g/, as in *rob, bed, big*, are sometimes realized as ... unreleased voiceless plosives [p̚, t̚, k̚]” but this “does not usually lead to loss of contrast with [p, t, k] ... because of allophonic effects on preceding vowels, thus for example *bet* /bet/ [bet] vs. *bed* /bed/ [bɛə̃t̚]” (Wells 1982: 558, based on Wolfram 1969: 102; see also Harris 1994: 210, Wolfram & Schilling-Estes 1998: 171). In spite of Edwards’s (2004: 388) claim that this feature of AAVE is not shared with the “Southern white vernacular dialect”, Pederson (2001: 279) provides the example of “devoiced /d/ in *hand*” for parts of Louisiana, but this may well be a lexical-incidental case which does not lead to neutralisation. Finally, according to Hartman (1985: lvii), “a partially devoiced variant [d̚] also occurs occasionally” in different American regional accents, but it occurs “especially in areas of foreign-language settlement”. Moreover, “variable post-sonorant /d/ devoicing ... after /l/ and /n/” has also been attested for Newfoundland English (Clarke 2004: 379). Again, it is unclear if this devoiced realisation of /d/ ever leads to actual neutralisation of the fortis/lenis contrast. This would, however, appear to be the case in the “English usage of the Pennsylvania Dutch”, who reportedly pronounce *made* as *mate* (McArthur 2002: 179; see also Anderson 2001, 2002), but no speakers of this variety of English have identified themselves as such in the experiment. In short, it is unlikely that any of these peripheral phenomena will make native speakers of any relevant variety of North American English significantly more lenient towards the substitution of word-final /d/ by /t/ commonly heard in Dutch English.

4.4.2 BAT

In North America, the opposition between /æ/ and /ɛ/ is normally firmly maintained.⁶ However, /æ/ and /ɛ/ may have a different distribution in the speech of both Spanish bilinguals and monolingual speakers of Chicano English, as a result of which *bat* may be pronounced as /bet/ while *bed* is realised as /bæd/ (Penfield 1985: 45; see also Penfield 1985: 36, Sawyer 1971: 378, Tottie 2002: 229). Another very common exception is the pre-rhotic “*merry-marry* merger” found, according to Trudgill & Hannah (2004: 44), in the Western, Midland and Northern varieties of General American. Pederson (2001: 272, 285)

⁶ As in Wells (1982), the symbol *e* has been used for the vowel in RP DRESS, and *ɛ* for the DRESS vowel in GA. The symbol *e* has been employed for any front open-mid realisations close to CV3. Where any American accent researchers have used the symbol *e* instead of *ɛ* (e.g. Labov 1966, Labov 1991, Pederson 2001), this has been retained.

also associates this merger with the Midland, the North and Northwest; according to Wells (1982: 480), it is found in “western New England and upstate New York ... and the middle and far west” (see also Gordon 2004b: 344). The contrast is retained, amongst others, by Canadians (Brinton & Fee 2001: 429) – although according to Boberg (2004: 357), this is only true of speakers from Montreal. Other than in pre-rhotic position, /æ/ does not normally merge with /ɛ/; an exception found in some Southern and South Midland speech (“/ɛ/ for /æ/ in *chance*”) is provided by Pederson (2001: 277; see also Thomas 2004: 308). But since the phonological context of these examples is very different from *bat* and *bet*, they are unlikely to bias any North American judges.

Realisations such as [ɛ] for /æ/ are equally infrequent: examples cited by Hartman (1985: xlvii) are “*that* and *ask* in New York City”, which may be “attributed to the influence of Eastern European Jews”. Labov (1966: 317) also notes that “Yiddish accents in English seem to favor the use of /ɛ/ for /æ/”. At the same time, realisations such as [ɛə, eə, ēə] and, in the “Eastern United States”, [iə] are beginning to be increasingly common (Wells 1982: 477). This phenomenon, termed “BATH raising” by Wells, may occur in various phonetic environments, including /_d/ and, less frequently, in /_t/, depending on the speaker’s accent and location and on the speech style used (Wells 1982: 477–478). Even by 1966, Labov (1966: 51) had noted that some New Yorkers “regularly use [ɛ:ə]” before “voiced stops”, but he also pointed out that before a “voiceless stop”, as in the present example, it is “almost always ... a short, checked vowel [æ].” More recently, Gordon (2004a: 285) has also observed that in “New York City, and elsewhere in the Mid-Atlantic region, the historical ‘short a’ vowel class is split into two phonemes”, which he labels “lax /æ/” and “tense /æə/”. The former “occurs consistently before voiceless stops”, while the latter may be “distinguished from the lax phoneme by lengthening and raising”, sometimes even to [ɛə], especially “among speakers from the lower end of the socioeconomic hierarchy” (Gordon 2004a: 285–286). Similar patterns are to be observed in Philadelphia (Gordon 2004a: 290–291) and Cincinnati (Gordon 2004b: 348). In other words, some speakers from New York, Philadelphia and elsewhere may pronounce *bath* with [ɛə] or [ɛ:ə], but they will still pronounce *bat* with [æ].

Be that as it may, these [ɛə, eə, ēə] realisations, which have already become part of “the British stereotype of an American accent” (Wells 1982: 477), may be re-interpreted by Dutch learners as [ɛ:] or even /ɛ/, possibly reinforcing their tendency to confuse English /æ/ with /ɛ/ (see, for instance, Collins & Mees 2003b: 94). Whether or not these Dutch realisations of /æ/ as [ɛ:] or /ɛ/ will in turn be perceived by any North American judges as allophones of the original /æ/ is a moot point. This is more likely to be true of those respondents who are either familiar with raised realisations of /æ/ or who produce these realisations themselves. The latter would include speakers involved in the Northern Cities Chain Shift, which Labov (1991: 14) locates in “western New England, New York State, the Northern Tier of counties in Pennsylvania, northern Ohio, Indiana and Illinois, Michigan, Wisconsin, and

a less-well defined are extending westward". In addition to raised /æ/, such speakers may also have lowered or backed /e/, resulting in either [æ̠] or [ʌ] (Labov 1991: 16–18).

It should be noted that the area identified by Labov as involved in the Northern Cities Chain Shift may in fact include judges who have labelled themselves as being from the Midwest and the Northeast. As has been demonstrated by Lance (1999), Americans have widely different mental maps of such elusive American dialect areas as "Midwestern" and "Northeastern". Depending to some extent on their regional provenance, they may well perceive some of the areas identified by Labov as belonging either to the Midwest, to the Northeast or to the North (Lance 1999: 284–296).

According to Labov (1991: 19), the Chain Shift may cause confusion for "listeners from another dialect area" who may perceive a Northern Cities "Ann as Ian" and "bet as bat or but". Conceivably, it could also make it easier for speakers involved in the shift to re-interpret foreign speakers' [ɛ:] or /ɛ/ as allophones of /æ/. As a result, /æ ~ ɛ/ confusion has been labelled as occurring in a minority of speakers from the "Northern/Northern Cities" area only. Speakers from New York City have not been included in this, even though, at least according to Trudgill & Hannah (2004: 47), New York City does take part in the Chain Shift. This decision is motivated by Labov's (1966: 51) earlier findings (confirmed in Gordon 2004a) that New Yorkers do not tend to raise /æ/ before a "voiceless stop".

4.4.3 VAN

According to Hartman (1985: lvii), there is a "a partially devoiced variant" of /v/, [v̥], which "occurs most often in areas of heavy German settlement", but it is unclear if this leads to actual /f ~ v/ confusion. In any event, the latter is not attested anywhere else in North America – other than as a "relic pattern occurring in scattered isolated communities" (Hartman 1985: liii). In word-final position, however, "devoicing of [v] to [f]" is found in Chicano English (Penfield 1985: 36 as discussed in Tottie 2002: 228). The related phenomenon of /f/ voicing may still be heard in some older speakers from Newfoundland, both according to Trudgill & Hannah (2004: 50) and to Clarke (2004: 380), but it may be difficult to predict its effect on any judges (especially since no listeners from Newfoundland took part in the experiment).

4.4.4 WINE

According to Hartman (1985: liii), interchanging /w ~ v/ is a "characteristic both of English Cockney speech and of creolized varieties of English" especially "among Blacks in the South Carolina-Georgia Low Country". Since it has actually been obsolete in London English for at least two centuries (see, for instance Wells 1982: 332–333), it is tempting to think that the same may be true for the relevant creoles. At any rate, no mention of it is made in Edwards's (2004) overview of the phonology of AAVE. But whether or not it is obsolete, the interchange of /w ~ v/ may well be related to the "sporadic pronunciation of

/v/ and /w/ as [β]” found in Gullah (Mufwene 2001: 297; see also Weldon 2004: 401). It is, apparently, also to be heard in the English of the Pennsylvania Dutch (McArthur 2002: 179). None of the judges, however, identified themselves as speakers of any of these varieties.

4.4.5 THIN, AUTHOR, BOTH, THAT, WEATHER, BREATHE

According to Wells (1982: 553), TH-fronting “in word-final position ... occurs only in lower-class, and particularly in black, speech” in the American South.⁷ Thomas (2004: 319–320) confirms that it is “much rarer” in the speech of rural Southern whites “than in African American speech”. Apart from TH-fronting in syllable codas, Wolfram & Schilling-Estes (1998: 324) also provide intervocalic examples heard in African American Vernacular English, such as “*efer* for *ether*” and “*brover* for *brother*”. In his overview of the phonology of AAVE, Edwards (2004: 388) also lists “[bæf]” and “[mʌvə]” as possible realisations of *bath* and *mother*. According to Hartman (1985: liii), such alternations” occur “occasionally, especially among conservative Black speakers”. TH-fronting has also been attested for Newfoundland English (Clarke 2004: 376). No judges identified themselves as speakers of any of these varieties, but if they had, their accents – to the extent that they use TH-fronting *as opposed to* TH-stopping – would have been labelled as different from Dutch English. This is simply because, as in the case of other varieties of English, the effect of TH-fronting on any respondents asked to evaluate samples of TH-stopping is very difficult to establish. (See also 4.2.5.)

Unlike TH-fronting, the phenomenon of TH-stopping appears to be quite common in North America. According to Hartman (1985: liii), it is

fairly widespread especially in northern urban areas (where it is often characteristic of working class speech), in the South (especially among Blacks), and in areas such as the Upper Midwest and the Southwest that have had dense settlement by foreign-language speakers.

But he also observes that

[s]uch cities as Milwaukee, Chicago, Cleveland, Buffalo, New York, Philadelphia and Baltimore, for example, have numerous speakers who use /t/ and /d/ for /θ/ and /ð/, a common substitution in foreign-accented English. Outside the South, however, these forms are not common in non-urban areas, even those heavily settled by non-English speakers (Hartman 1985: xlvi).

Whether or not such substitutions involve initial, medial and final /θ/ and /ð/ all to the same extent is unclear, as is the question of whether or not TH-stopping is actually found in northern rural areas with “dense settlement by foreign-language speakers”, such as “the Upper Midwest”. In any case, Gordon (2004a:

⁷ For definitions of TH-fronting and TH-stopping, see 4.2.5.

298) also reports TH-stopping in “urban speakers” of the Inland North. In addition, individual judges, however, make observations such as the following:

The **th** pronounced as **t** [as in authority] does sound native to the mid-northern states of the US such as Minnesota and Wisconsin – but it is an [e]ffect that most Americans find comical. It is often exaggerated in humorous skits about Canadians and Minnesotans (Subject 653).

or produce claims such as these: “I live in Wisconsin, and some people here actually use ‘d’ in place of ‘th’”. This is most common very far north[.] ‘Hey there!’ becomes ‘Hey der!’, for example” (Subject 702). It also appears to be a well-known fact that some Newfoundland speakers have TH-stopping (Clarke 2004: 376, Kirwin 2001: 449, Wells 1982: 498, 500), but since there are no judges from that area, they may be excluded from consideration. Realisations of /θ/ and /ð/ as [t] and [d] or [tθ] and [dð] (Schneider 2004b: 1084) have also been attested for AAVE (Edwards 2004: 388), Cajun Vernacular English (Dubois & Horvath 2004: 411), Gullah (Weldon 2004: 402), Philadelphia English (Gordon 2004b: 293) and as occurring sporadically in the speech of rural Southern whites (Thomas 2004: 319). However, no respondents indicated that they were speakers of any of these varieties.

Not only does New York City have clearly documented incidence of TH-stopping involving both /θ/ and /ð/, but unlike Newfoundland and other areas it also provided judges that have taken part in this experiment. As in Newfoundland (Wells 1982: 498, 500), it is often dental stops such as [t̪] or [d̪] that are used, at least by some New Yorkers, for /θ/ and /ð/, so that oppositions such as *thin* ~ *tin* or *that* ~ *dat* tend to be preserved (Wells 1982: 515–517). While some descriptions of New York English state in general terms, as Gordon (2004a: 288) does, that “/θ/ and /ð/ are often realised as stops ... or affricates”, others distinguish clearly between the initial, medial or final position of these sounds. For instance, Wells (1982: 516) observes that the “/ð ~ d/ opposition seems to be lost rather more readily”, at least, to some extent, in initial position, and in “one or two words in which the /ð/ is not initial, e.g. *other*... . But it would not be usual for *southern* to be pronounced identically with *sudden*, or *breathe* with *breed*”. Wolfram & Schilling-Estes (1998: 325), too, note that “[s]ome restricted Anglo varieties use a stop *d* for intervocalic voiced *th* as in *oder* for *other* or *broder* for *brother*, but this pattern is much less common than the use of a stop for *th* in word-initial position”. This would suggest that when TH-stopping occurs word-medially or finally in “Anglo” or “Euro-American” accents such as that of New York City, it is so rare as to be unlikely to influence any listeners.

If TH-stopping is much more common word-initially, it is also more frequently employed with /ð/ than with /θ/. In Black English, for instance, word-initial TH-stopping is in fact largely restricted to /ð/ (Wells 1982: 558) – although Edwards (2004: 388) also cites examples such as “[tɪŋ]” and “[tɛnt]” for *thing* and *tenth*. According to Wolfram & Schilling-Estes (1998: 324), initial alternation between /θ/ and /t/ “tends to be most characteristic of selected Anglo-

and second-language-influenced varieties”, whereas the equivalent alternation between /ð/ and /d/ “is spread across the full spectrum of vernacular varieties”. In Chicano English, however, it appears to involve both /ð/ and /θ/, and is particularly common with Chicano speakers from the “lower socio-economic class” (Penfield 1985: 42–43; see also Tottie 2002: 228).

If TH-stopping is so very widespread (see also Schneider 2004b: 1084), it could be argued, as Pederson (2001: 260) does, that it is a “social” rather than a regional “marker” which helps to “distinguish American dialects ... irrespective of their geographic provinces”. Especially initial /ð ~ d/ alternation is a “social stereotype” (Wolfram & Schilling-Estes 1998: 161) which “may even lead to the stigmatization of speakers of ‘stupid’ and ‘uneducated’” (Wolfram & Schilling-Estes 1998: 75; see also Penfield 1985: 43). Incidentally, judges who are familiar with such realisations do not necessarily have to be biased against them. As one of the respondents pointed out: “[M]y judgements about seriousness of **t/d** substitutions for **th** are influenced by the local dialects I hear of Cajun English and Black English, and I think I judge an error as more serious if it doesn’t occur locally! [S]orry” (Subject 432).

The pervasiveness of word-initial /ð/ stopping as a social marker throughout North America, extending beyond accents that are stereotypically credited with the phenomenon, would suggest that it affects all North American judges who were asked to evaluate the phenomenon in Dutch English – with the possible exception of GA speakers. It would be difficult to prove that this extends to judgements about /ð/ stopping in other environments (intervocalic or word-final) – although this is not unlikely. In the absence of unambiguous data, it will be assumed that while word-initial /θ/ stopping is found in New York City, its presence in other relevant environments and in relevant accents has been documented too infrequently to warrant inclusion.

4.4.6 OFF

In most varieties of English, substitution of /v/ by /f/ is normally only found in the context of assimilation with a following consonant. This is normally restricted to “unstressed syllables ... with final inflexional /d/ and /z/ , and also with grammatical items such as *as* and *of*, and auxiliary verbs” (Collins & Mees 2003b: 210) and clitics (Selkirk 1972: 186). Substitution of /v/ by /f/ before a vowel, however, as in this token, has not been attested for North American English. A possible exception to this may be found in what Hartman (1985: lvii) refers to as “areas of heavy German settlement”.

4.4.7 RED

The weakly voiced uvular fricative [ʁ] produced by the GA actor has not been attested for any of the relevant accents, although it has been attested in “isolated pockets of Newfoundland” (Hickey 2004a: 79; see also Clarke 2004: 377). Nevertheless, it is quite likely to be a feature of some immigrant speech, as [ʁ] is to be heard in a wide variety of languages, including “Dutch, Norwegian and Swedish” and “is now standard in French, German and Danish” (Trudgill 1984a:

56). According to Saciuk (1989), it is also common in Puerto Rican Spanish. Two judges referred to it as “Israeli” or “Hebraic”. As Trudgill (1984a: 57) points out, uvular-*r* does indeed occur in “some varieties of Afrikaans, Hebrew and Canadian French”. None of the judges, however, identified themselves as speakers of any of these languages or varieties. In fact, they repeatedly described this realisation of /r/ as “foreign”.

4.4.8 ICE

The realisation [a:i] has not been attested for any of the relevant North American accents (see Hartman 1985: lvi).

4.4.9 TIE

There is no evidence to indicate that aspiration of initial [t] is either absent, hardly noticeable or in any way obsolescent in the relevant varieties of native North American English (see, for instance, Collins & Mees 1993: 14, Hartman 1985: lvii, Gimson & Cruttenden 1994: 151). Interestingly, however, it has been reported for Cajun Vernacular English (Dubois & Horvath 2004: 411) and Gullah (Weldon 2004: 400).

4.4.10 DEAD

In none of the relevant North American accents is /d/ ever realised as [ʔ]. Hartman (1985: lvii), for instance, lists [ʔ] as a variant of /t/, not of /d/. Similarly, the examples of North American glottalling as provided by Wells (1982: 501, 515, 553) all involve realisations of /t/, the one exception being Black English or African American Vernacular English, where “single final /b, d, g/, as in *rob*, *bed*, *big*, are sometimes realized as a glottal plosive [ʔ]” (Wolfram 1969: 102, as discussed in Wells 1982: 558). Edwards (2004: 388) also cites “[bæt]” and “[bæʔ]” as variant AAVE realisations of *bad*. However, none of the judges described themselves as speakers of Black English, although one respondent stated that this realisation of /d/ in *dead* is “[v]ery common in NY” and “[a]lso common among African Americans” (Subject 467). It should be stated, though, that this participant identified the error in question as “[u]nvoiced /t/ for /d/” – as did other judges who described the error explicitly. Since this respondent claims to be a speaker of “Standard American”, it would be difficult to establish whether or not these observations are based on intimate knowledge of either New York or Black English. On the basis of these comments, it has not been possible to categorise any of the relevant varieties as being similar to Dutch English in this respect – or in any other (such as BED).

4.4.11 FILM

Pronunciations such as [fɪləm] appear to be widespread in the United States, but are frequently stigmatised. Tellingly included in Cassidy’s (1985b: xxxvii) “Language Changes Especially Common in American Folk Speech” (together with *ellum* for *elm*), they are described by Mencken (1952: 99–100) variously as “probably of American origin” (but, according to a footnote, also possibly of

Dutch extraction), as a linguistic feature of what he terms the “movie Zion” of Hollywood, and, again in a note, as a humorous caricature. According to *DARE* (*Dictionary of American regional English*), the pronunciation /fɪləm/ is “also freq[uent]” as a variant of /film/ and has been observed as far afield as Arizona, New York, Texas, Kentucky and Pennsylvania (Cassidy 1991: 414). There is even more data on the similar pronunciation of /ɛləm/ for *elm*; according to *DARE* this is “widespread but somewhat more freq[uent] in the] S[ou]th, S[ou]th Midl[and], esp[ecially] freq[uent] among men and among rural Inf[ormant]s, somewhat old-fash[ioned]” (Cassidy 1991: 290). It is also supposed to be a notable feature of New York English (Branford 1994: 486). Judges from different parts of the United States repeatedly confirmed the stigma attached to this pronunciation. As a result, it has been labelled as occurring in a minority of speakers of regional American accents – with the obvious exception of GA. There is no information, either from the judges or from existing literature, on the situation in Canada. The exception to this is the speech of conservative Newfoundlanders (see Clarke 2004: 379), but none of the respondents have identified themselves as hailing from that province.

4.4.12 CAR

Even though some North American judges – notably those speaking rhotic accents – may have reservations about non-rhotic pronunciations of *car*, it is impossible for them to object to the particular manner and place of articulation of an absent consonant. This means that for the North American version of the experiment, there is no need to distinguish between judges’ possible objections to the presence of /r/ and their evaluations of the realisation of this phoneme. Since most varieties of North American English are rhotic, speakers of these accents will pronounce *car* with a clearly audible /r/. As far as the pronunciation of this token is concerned, they are similar to the majority of Dutch learners of American English, but different from this particular GA actor. It is only speakers of non-rhotic (or variably rhotic) accents that are likely to more lenient towards an r-less pronunciation of *car*. These accents are generally found in “most of the South and along much of the East Coast” (Hartman 1985: lviii). Interestingly, the judges who volunteered additional comments associated r-lessness only with the East Coast (or Britain), rather than with the South.

In fact, the two most widespread non-rhotic varieties on the East Coast, New York and Eastern New England, are subject to considerable pressure from GA to conform to the rhotic prestige norm (Wells 1982: 506, 520–521). It is clear from Labov’s 1966 study that New York English is variably rhotic. The variability is also evidenced by Wells’s (1982: 507) description of the accent as “basically non-rhotic”. This suggests that non-rhoticism is a feature found in the majority of speakers of New York English. While, according to Gordon (2004a: 288), “/r/ continues to divide New Yorkers along class lines ... the trend toward rhoticity appears to be progressing”. At the same time, what Wells (1982: 520) refers to as eastern New England’s “return to rhoticity” seems particularly apparent in “the speech of younger speakers” (Trudgill & Hannah 2002: 46).

While this may suggest that on the eastern seaboard the majority of speakers have retained their non-rhoticity, it does not mean that any respondents identifying themselves as hailing from New England as a whole are most likely to be non-rhotic. The same goes for those from the Northeast.

In the South, the distribution of rhotic and non-rhotic accents is also quite complex and subject to change (Thomas 2004: 317–318). According to Trudgill & Hannah (2002: 40–42), the “Lower South” is “generally ... non-rhotic”, whereas the “Inland Southern accents” are “typically, if sometimes variably, rhotic”. Wolfram & Schilling-Estes (1998: 160) point out that “the valuation of *r*-less speech has changed over the decades, and today it is working-class rural groups in the South who are most characteristically *r*-less rather than urban upper-class speakers”, unless the latter belong to the older generation, in which case they would tend to be non-rhotic. Thomas (2004: 318) also refers to a “dramatic increase in rhoticity” in white Southerners, while Tillery & Bailey (2004: 334) state that in last 25 years, “the expansion of rhotic variants has been so extensive among white Southerners that non-rhotic forms are now primarily associated with African Americans”. In any event, Wells (1982: 542) makes clear that even “those southern accents which are non-rhotic are often only variably non-rhotic” (see also Thomas 2004: 317). Such a claim may also be true of African American Vernacular English (Wells 1982: 543) and Gullah (Weldon 2004: 402). This would suggest that full non-rhoticity is only found in a minority of Southern speakers. It is also a feature of Cajun Vernacular English (Dubois & Horvath 2004: 412).

4.4.13 HOT_TEA

Degemination across word boundaries has not been attested in the native English of any North Americans. According to Wells (1982: 501, 552–553, 558), there are various phenomena that involve cluster reduction, notably in Newfoundland, the American South and in Black English, but none of these apply to geminated consonants or across word boundaries. (See also Schneider 2004b: 1087.)

4.4.14 NEW

New may be pronounced in North America variously as /nu/, /nju/, /niu/ or /nrɪ/ – with realisations such as [ɪɹ] or [jɹ] in the interior of the South, [ɹ] in the South Midland, and [iu] or [iɹ], “which prevail to the east and south” (Pederson 2001: 280, 273). The predominance of /nu/ in GA has also affected accents where the traditional prestige norm was /nju/, as in Canadian English. In the latter accent, both pronunciations are becoming equally prestigious – in spite of the fact that some Canadian media display a tokenist adherence to pronouncing a glide in “their most frequent and undoubtedly most salient lexical item”, *news* (Clarke 1993b: 87, 104). According to Boberg (2004: 356), “younger Canadians now ... delete the glide in words like *news* and *student* pretty much to the same extent and in the same environments as most Americans do”. Wells (1982: 489) notes that the GA “preference for ... /nu/ in ... *new* ... is, however, subject to

pressure from schoolteachers who often prescribe ... /nju/ as correct. The tendency of midwestern radio announcers to hypercorrections such as /nju/ *noon* is notorious”.

But it is not only in GA and in accents strongly affected by GA that considerable variation can be found. While Hartman (1985: lii) states that “/ju/ also occurs frequently on the Atlantic Coast from Delaware and Maryland south into Florida, then westward into Arkansas and eastern Texas, as well as in the San Francisco area and Hawaii”, with “scattered usage in the rural West” (lii), the *DARE* entry for *new* calls both pronunciations “widespread”, with /nu/ more common in the North and North Midland and /nju/ “more freq[uent]” in the South and South Midland (Cassidy & Houston Hall 1996: 780). In addition, Thomas (1947: 156–57) notes that [ju] may also be heard in Eastern New England and New York City.

Interestingly, Lippi-Green (1997: 36) describes /nu ~ nju/ variation as a sociolinguistic variable “which has more social currency in the south than it does in the north”. Thomas (2004: 319) even claims that yod-retention has in fact “persisted in the South longer than in any other part of the United States (though it still appears elsewhere as an affectation)”, but notes that after “World War II ... a steady movement towards loss of [j] in the South has occurred”, which “has been slower in common words” (such as *new*) “than in infrequent words”.

According to Wells (1982: 539, 504), “the falling diphthong /iu/ (or /ɪu/)” is found in speakers from New England, the South and from New York City. In the South, for instance, speakers may pronounce *tune* variously as /tiun/, /tun/, or /tjun/ (Wells 1982: 539). Since the difference between /iu/ and /ju/ is perceptively very small, the GA actor’s pronunciation of this token is unlikely to affect judges who have /iu/ rather than /ju/.

All this would suggest that there are very few areas in North America where /ju/ or /iu/ does not exist as a minority pronunciation. After all, there are many GA speakers who pronounce *new* as /nju/. As a result, all relevant North American accents have been labelled as being similar to Dutch English in this respect, for at least a minority of speakers.

4.4.15 IMAGIN

There is not enough reliable data to suggest that [ɪmædʒə'neɪtɪv] is an acceptable minority pronunciation in North America. To start with, Kenyon & Knott (1953: 240) only list [ɪ'mædʒə'netɪv] – with [ɪ'mædʒə'nəɪv] as a variant pronunciation (see also Kenyon & Knott 1953: 31). In addition, the *Longman pronunciation dictionary* gives [ɪ'mædʒə'neɪtɪv] as an alternative pronunciation for GA (Wells 2000: 381), but not with the stress on the penultimate, as in the experiment. In fact, none of the judges recognise the latter as an acceptable regional variation. Admittedly, it is actually recorded as an alternative pronunciation in the *American heritage dictionary*, but since this is not a pronouncing dictionary, its

authority may be questioned.⁸ In actual fact, it would appear to be a common pattern in American English for the first syllable of the sequence “-ative” to lose its stress in the environment of a preceding vowel plus sonorant (see Nanni 1977: 757).

For the sake of completeness, it may be interesting to note that one Australian judge of the RP version of the experiment stated that it “sounds odd to native speaker to have [the] fourth syllable stressed (except for perhaps some regional varieties of American English)”. Interestingly though, this respondent then went on to insist that “[n]o stress” should be given to any syllable” (Subject 999).

4.4.16 PERFECT

No differences from other varieties of English have been attested. Although the entry in Kenyon & Knott (1953: 324) mentioned E[astern] and S[outhern] realisations of the verb with the stress on the final syllable, for General American this was glossed as “now less freq[uent]”. Since this dictionary appeared more than half a century ago, it is quite likely that such information is outdated. Support for this is found in there being no mention in either the *Longman pronunciation dictionary* (Wells 2000) or the latest *EPD* (Jones 2003) of the possibility in American English of the verb exhibiting anything other than final stress.

4.4.17 TO_WALES, THAT_THA, WOULD_ON

Overuse of Strong Forms has not been found in any of the relevant varieties of North American English.

4.4.18 SECONDAR

There is insufficient evidence to assume that [ˈsekəndrɪ] is a common alternative of [ˈsekənderi] in any relevant variety of North American English. To start with, the *Longman pronunciation dictionary* only gives [ˈsekənderi] or [ˈseknderi] for GA (Wells 2000: 685). Similarly, Kenyon & Knott (1953: 380) only list “[ˈsekən,deri]”. *DARE* does not have an entry for *secondary*, but *secretary* appears as “usu[ally] /ˈsekrətɜːrɪ/”. All the regional variations for *secretary* recorded in *DARE* have the DRESS vowel in the penultimate (Houston Hall 2002: 836). Bauer (2002: 81), who also discusses the pronunciation of *secretary* in various accents of English, confirms that in GA or Canadian English, the penultimate is only ever pronounced with the DRESS vowel, as is the case with *monastery*. According to Brinton (2001: 430), “[t]he retention of secondary stress in words ending in -ory, -ary, and -ery ... is standard in Canadian English as it is in American English, thus being distinguished from British English”.

⁸ See entry for “imaginative” in the *American heritage dictionary*, 2nd college edition (1982: 524), or in online version of 4th edition at www.bartleby.com/61/27/I0042700.html (accessed 16 May 2006).

Over half a century ago, Mencken (1949: 267, 1952: 32) had already derided and stigmatised the elision or weakening of the penultimate as a “Briticism” or as a “mutilated form”: “In *secretary* what the Englishman does is to get rid of a syllable altogether, so that the word becomes, to American ears, *secetry*; the American himself almost always gives it its lawful four” (Mencken 1952: 4). Most judges concur that [sekəndrɪ] is a British rather than an American pronunciation. In the few cases where it is described as an alternative regional pronunciation, the respondents concerned do not appear to be speakers of these regional accents themselves. For instance, a West Coast listener located this feature in New England (Subject 696), whereas a participant from the American South termed it Canadian (Subject 944). As one judge put it: “*Secondary* is pronounced as many Eastern Americans might pronounce it. In the Midwest and West, we tend to enunciate syllables ...” (Subject 653).

4.4.19 TELL

According to Wells (1982: 490), GA “/l/ tends to be rather dark” and is velarised in final position. This would also appear to be true of final /l/ in other US and Canadian accents except Newfoundland (Wells 1982: 495, 498). Trudgill & Hannah (2002: 39) confirm that in “most” of these “varieties, /l/ is fairly dark in all positions”.

It is true, as Wells puts it, that the “phoneme /l/ itself exhibits greater allophonic differences in the south than in other parts of North America” (Wells 1982: 550). Since, as Wells points out, the difference in the realisation of /l/ is “particularly noticeable” in intervocalic position (Wells 1982: 490), this will not bias Southern judges unduly. It is unclear whether or not the Southern potential realisation of dark /l/ as a “velar lateral, which may be symbolized [L]” has an equivalent in Dutch English (Wells 1982: 551). In the absence of any detailed information on this, it will be assumed that it is also unlikely to affect any Southern respondents.

Collins & Mees (1993: 34) note that “Dutch dark [ɫ] is produced with the root of the tongue drawn back ... making it sound too ‘hollow’ to an American ear”; this “may give the impression that /l/ is missing altogether” (see also 3.5.20). This would suggest that instead of [ɫ], some Americans may well interpret this as L-dropping, especially if this phenomenon exists in their own variety, as it does in the South and in Black English. Since, according to Wells (1982: 550–551), Southern L-dropping, which occurs only before a “labial or velar”, is “quite strongly stigmatized”, it is extremely unlikely to render Southern judges more lenient towards L-dropping in other environments such as they may perceive in Dutch speakers. Omission of final /l/ is also common feature of African American Vernacular English, as in *toll* (Mufwene 2001: 296) or [rɔ] for *roll* (Edwards 2004: 388), but no self-styled speakers of Black English took part in the experiment.

According to Wells (1982: 517–518), L-vocalisation “is quite common in New York, though not on the scale found in the English or the American south”, is found in environments such as “*sell*”, and does not appear to be “confined to

uncultivated speech”. “The Inland South” is also mentioned by Gick (2002: 170) as a dialect “well known for particularly extreme vocalization of /l/”. Tillery & Bailey (2004: 334) also confirm that in the American South, “post-vocalic /l/ is frequently vocalized”, both in urban and “rural varieties”. In view of the fact that L-vocalisation is common, but not universal, in New York and Southern English, these two varieties have been categorised as being similar to Dutch English in this respect for a minority of speakers. L-vocalisation is also very common in Philadelphia (Ash 1982: 162, as quoted in Gick 2002: 169; see also Gordon 2004a: 293), but no judges from that area identified themselves as such. The same is true for Newfoundland English (Clarke 2004: 377) and AAVE (Edwards 2004: 388). In addition, L-vocalisation occurs in “traditional Midland areas” and is “reported to be a characteristic of Pittsburgh speech” (Gordon 2004b: 342). There may have been respondents from these areas who actually identified themselves differently (for instance, as “Midwestern” or “Standard American”), but in the absence of proof this has been excluded from further consideration.

4.4.20 COLOUR

In Newfoundland speech, “the mid-central nucleus of *cut*, *slub*, *dull* is a backed vowel, often rounded, in effect approaching the area which in some other dialects is occupied by [ɔ] (Kirwin 2001: 449; see also Wells 1982: 498). Clarke (2004: 371) states that for “the Irish Avalon ... the vowel is best represented as [ɔ̃]”. No self-styled Newfoundlanders, however, took part in the experiment. In addition, there are examples of conditioned /ʌ ~ ɔ/ mergers, such as pre-velar /ɔ/ in *donkey* or *honk*, which some speakers in Pennsylvania, New Jersey, and New York may pronounce as /ʌ/ (Thomas 1947: 151, 157). These appear to be infrequent, recessive and not spreading to other environments. More recently, speakers taking part in the Northern Cities Chain Shift have been observed to pronounce buses as [bʊsɪz] (Labov 1991: 18). The geographical spread of this pronunciation has been detailed in TELSUR map IN-3, and includes younger speakers from cities such as Milwaukee, Chicago, Detroit and Buffalo (Labov *et al.* 2005b). In all likelihood, this will only make a minority of speakers from the North, the Northeast and the Midwest (even in those cases where these labels refer to the Northern Cities area; see BAT) more lenient towards Dutch speakers of English who pronounce *colour* identically with *collar*.

4.4.21 STOOD

There is no evidence to suggest that /ʊ/ in *stood* is realised as /u/ (the GA equivalent of RP /u:/) in any native variety of North American English. Admittedly, there are quite a few examples of /ʊ ~ u/ mergers in American accents, but these normally take place before a sibilant or /l/. According to Trudgill & Hannah (2002: 45), for instance, even “educated speakers” from what they term the “Midland” area may pronounce /ʊ/ “before the fricatives /ʃ/ or /ʒ/” as /u/, as in *push*. Hartman (1985: li) also mentions the “South Midland” and “the rural West” as areas where words like *bush*, *push*, *butcher* may be

pronounced with /u/. Words such as *cool*, *fool*, *pool*, where the vowel precedes /l/, are also occasionally pronounced with /ʊ/, particularly by “younger speakers in the West” (Hartman 1985: li). Wolfram & Schilling-Estes (1998: 71) cite *pull* and *pool* as examples of a similar merger in “Texas and the South”. The phenomenon is also discussed in some detail in the online *Atlas of North American English*: map 6 in particular shows the extent of the merger of /u/ and /uw/ before /l/ (Labov *et al.* 2005a). An overview of such “pre-L mergers” is provided in Gordon (2004b: 344–345), who points out that some speakers involved in these sound change mergers “may perceive no contrast between the sounds even when they consistently produce a distinction phonetically” (345).

Since none of these environments are similar to that in *stood*, North American listeners are unlikely to be affected by these mergers when assessing the GA actor’s realisation of this word. This would also appear to be true in those few cases where words of a “limited set” like *coop*, *root*, *roof*, *hooves*, *room* are pronounced, at least by some speakers, scattered across the US, with /ʊ/ (Hartman 1985: lii).

4.4.22 INT1, INT2, INT3

Even though there are noticeable differences between North American and some other English accents in terms of intonation patterns (Pederson 2001: 261), this does not mean that North American judges are any more likely to evaluate Dutch intonation patterns more leniently than those speaking with a British, Irish or Antipodean accent. (See 4.2.22.)

4.5 Accent similarity codes for the GA version of the experiment

On the basis of the discussion in 4.4, similarity codes have been assigned to all severity judgements in any one minor accent group taking the GA version of the experiment. This is shown in Table 4.3. The “SIMMIN” and “SIMMAJ” codes are in bold.

Table 4.3. Accent similarity codes for all minor accent groups taking the GA version of the experiment (by token).

	<i>US/GA</i>	<i>CDN</i>	<i>US/EC</i>	<i>US/WS</i>	<i>US/NC</i>
<i>BED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BAT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN
<i>VAN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WINE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>AUTHOR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

<i>BOTH</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>OFF</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT</i>	DIFMAJ	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>WEATHER</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BREATHE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>RED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>ICE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TIE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>DEAD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>FILM</i>	DIFMAJ	DIFMAJ	SIMMIN	SIMMIN	SIMMIN
<i>CAR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>HOT TEA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>NEW</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>IMAGIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>PERFECT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TO WALES</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT THA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>SECONDAR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WOULD ON</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TELL</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>COLOUR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	SIMMIN
<i>STOOD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT1</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT2</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT3</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

	<i>US/MW</i>	<i>US/NE</i>	<i>US/NY</i>	<i>US/SO</i>
<i>BED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BAT</i>	SIMMIN	SIMMIN	DIFMAJ	DIFMAJ
<i>VAN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WINE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THIN</i>	DIFMAJ	DIFMAJ	SIMMIN	DIFMAJ
<i>AUTHOR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BOTH</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>OFF</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>WEATHER</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>BREATHE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>RED</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>ICE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TIE</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>DEAD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

<i>FILM</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>CAR</i>	DIFMAJ	SIMMIN	SIMMAJ	SIMMIN
<i>HOT TEA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>NEW</i>	SIMMIN	SIMMIN	SIMMIN	SIMMIN
<i>IMAGIN</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>PERFECT</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TO WALES</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>THAT THA</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>SECONDAR</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>WOULD ON</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>TELL</i>	DIFMAJ	DIFMAJ	SIMMIN	SIMMIN
<i>COLOUR</i>	SIMMIN	SIMMIN	DIFMAJ	DIFMAJ
<i>STOOD</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT1</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT2</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ
<i>INT3</i>	DIFMAJ	DIFMAJ	DIFMAJ	DIFMAJ

4.6 Accent similarity: analysis and results

The analysis discussed below shows that if a pronunciation feature is likely to occur in the accent group of the respondents, it may often be assessed more leniently. This overall effect was found for all 20 relevant tokens when pooled. Nevertheless, the effect was only attested in a relatively small number of individual tokens (BOTH, THAT, WEATHER, CAR, NEW). Furthermore, one token (STOOD) was assessed significantly less leniently by respondents whose accents are likely to have a comparable pronunciation feature. In a number of other cases, accent similarity did not have a demonstrable effect at all. Even BAT and TIE were not judged significantly differently, even though in some accents, these pronunciations occur in a majority of speakers. Clearly, it is difficult to generalise the effects of accent similarity, and consideration should be given to judges' varying attitudes to different realisations. It is telling that respondents may sometimes even assess Dutch English realisations that are similar to their own just as severely as other native speakers – in some cases, at least, there is no evidence to suggest that they evaluate such pronunciations more leniently.

The MLwiN program was used to estimate, for all 20 relevant tokens put together, the average severity for all responses coded either “DIFMAJ”, “SIMMIN” or “SIMMAJ” (dependent on the expected absence or presence, in the accent group of the judge in question, of a realisation similar to that in the carrier sentence). If listeners are indeed more tolerant of pronunciations similar to those found in their own accents, one would expect a higher severity estimate for responses coded “DIFMAJ” than for those coded “SIMMIN” or “SIMMAJ”.

This is because they would be unlikely to judge an unfamiliar realisation more leniently. Analysis showed that this was in fact true of “DIFMAJ” and “SIMMIN”: the difference between the estimated average severity for responses coded “DIFMAJ” (1.872; s.e. 0.088) and “SIMMIN” (1.231; s.e. 0.094) was actually highly significant ($\chi^2 = 212.59$, $df = 1$). The estimated average severity for “SIMMAJ” (1.44; s.e. 0.183) was not significantly different from either “DIFMAJ” or “SIMMIN”, neither at $\alpha = .05$ nor at a less strict $\alpha = .10$. This result may be connected to the fact that the “SIMMAJ” code had only been assigned in a relatively small number of cases. This is because in most cases, there was no definite proof that a majority of speakers had realisations similar to the token in question.

As it is impossible to quantify the distinction between “similar for a majority of speakers” as opposed to “similar for a minority of speakers” very precisely, and as the estimates for “SIMMAJ” were not significantly different from the other two, it was decided to develop a new interpretation of the test results, one which recodes all responses formerly coded either “SIMMAJ” or “SIMMIN” as “SIM”, and all responses previously coded “DIFMAJ” as “DIF”. Severity estimates for these new variables were subsequently calculated using the MLwiN program. The difference between the lower estimate for “SIM” (1.243; s.e. 0.094) and the higher one for “DIF” (1.873; s.e. 0.018) was in fact highly significant ($\chi^2 = 213.36$, $df = 1$). Since these figures are based on all 20 relevant tokens combined, this means that there is a general tendency for pronunciations that are likely to be found in the accents of respondents to be judged more leniently.

A breakdown of the severity estimates for “DIF” and “SIM” by token, as in Table 4.4, reveals that only in the case of five tokens (BOTH, THAT, WEATHER, CAR, and NEW) were those for “SIM” significantly higher than those for “DIF”. Remarkably, STOOD had a significantly higher severity estimate for SIM than for DIF, so this token was in fact judged *less* leniently by those respondents who are likely to have a similar realisation in their own accents.

Firstly, accent similarity clearly does not necessarily affect the evaluation of all tokens. Nevertheless, it should be pointed out that the lack of significant differences for the remaining fourteen tokens could also be due to large sampling errors. Table 4.5 shows, for instance, that for BED, VAN and DEAD the number of responses coded “SIM” is actually lower than 10. However, this does not explain why there were no significant differences for COLOUR, FILM or RED. For these tokens, there are a great many responses coded “SIM”, so they are unlikely to be affected by large sampling errors.

Secondly, another striking result is that accent similarity may also cause respondents to judge a token more severely, as in the case of STOOD. It may be argued that this is restricted to this token only. But there are other examples from this experiment that appear to illustrate a tendency for respondents to judge pronunciations more severely if they are familiar with them.

Table 4.4. Estimated average severity: differences between responses coded “DIF” and “SIM” (for selected tokens). These are significant at $p < 5$ ($\chi^2 > 9.14$, $df = 1$), unless listed as “n.s.”.

Token	DIF	Standard Error	SIM	Standard Error	χ^2	
BED	2.449	0.189	2.521	0.606	0.01	n.s.
BAT	2.443	0.214	2.216	0.259	1.60	n.s.
VAN	2.622	0.209	2.657	1.534	0	n.s.
THIN	2.658	0.165	2.833	0.195	1.99	n.s.
AUTHOR	2.290	0.159	2.341	0.190	0.17	n.s.
BOTH	1.507	0.246	0.555	0.307	20.80	
THAT	0.904	0.231	1.548	0.247	22.53	
WEATHER	2.001	0.226	1.073	0.276	23.50	
BREATHE	1.617	0.276	1.827	0.336	0.89	n.s.
RED	2.651	0.190	2.373	0.218	3.85	n.s.
TIE	1.576	0.240	1.710	0.287	0.46	n.s.
DEAD	2.359	0.230	1.681	0.610	1.36	n.s.
FILM	2.244	0.211	1.989	0.226	4.06	n.s.
CAR	2.141	0.098	1.691	0.122	11.81	
NEW	1.173	0.125	-0.038	0.090	166.45	
THAT_THA	1.054	0.227	1.338	0.304	1.58	n.s.
SECONDARY	1.510	0.217	1.521	0.222	0.01	n.s.
TELL	1.20e-15	6.65e-09	-1.71e-15	5.70e-09	0	n.s.
COLOUR	2.377	0.195	2.400	0.221	0.03	n.s.
STOOD	1.977	0.232	2.600	0.266	12.13	

Table 4.5. Number of responses coded variously “DIF” or “SIM”, for selected tokens.

TOKEN	DIF	SIM
BED	538	7
BAT	454	91
VAN	544	1
THIN	463	82
AUTHOR	468	77
BOTH	468	77
THAT	324	221
WEATHER	456	89
BREATHE	468	77
RED	443	102

TOKEN	DIF	SIM
TIE	479	66
DEAD	537	8
FILM	349	196
CAR	365	180
NEW	262	283
THAT_THA	497	48
SECONDAR	369	176
TELL	392	153
COLOUR	414	131
STOOD	479	66

It was, for instance, suggested in 3.5.1 and 3.5.11 that many North Americans assess BED and DEAD more severely, not because they have these realisations themselves, but because they associate them with stigmatised varieties of American English, such as AAVE. In other words, such increased severity is likely to occur when a particular realisation is strongly stigmatised within a particular accent group.

In the case of STOOD, however, the only accent groups that have been coded “SIM” are GB/SC, GB/SG and IRL/N. These Scottish and Northern Irish judges are unlikely to object to [stʊd] for *stood*, since this appears to be one of the most common pronunciations in Scotland and Northern Ireland (see 4.2.21). Neither were there any comments from Scottish or Northern Irish participants that hinted at possible stigmatisation. However, there are a number of other possible reasons why they judged this token more severely. For instance, they may have objected to this realisation if they perceived the vowel as too long, i.e. in contravention of Aitken’s Law (Wells 1982: 400). According to this phonological rule, also known as the Scottish Vowel Length Rule, the vowel in *stood* is invariably short in Scottish English. Another option is that these respondents may have felt that it did not fit in with the RP accent of the speaker.

Thirdly, it is interesting that the tokens whose severity estimates were significantly divergent for “DIF” and “SIM” are a subset of those judged significantly differently in the RP and GA versions of the experiment (see 3.3.4.3). Of these, BOTH, THAT, WEATHER and CAR were evaluated more strictly by North American respondents, while NEW and STOOD were assessed less leniently by participants in the RP form. The accent similarity results suggest that the stricter North American assessment of a number of tokens, and the less lenient judgement of other tokens in the RP form, may not extend to judges who have similar features in their own accents. This is a useful corrective to the notion that all respondents in a particular version judge certain tokens more seriously across the board. For instance, while the data suggest that non-rhotacism is considered more serious in North America than rhotacism is in Britain, Ireland or the Antipodes, this may not apply to non-rhotic judges in America or to rhotic judges in Britain, Ireland or New Zealand. Non-native learners of English are well-advised to take this into consideration when deciding which pronunciation issues to prioritise.

Fourthly, however, reservations about uniformly severe assessments of errors do not appear to apply to such realisations as RED, FILM or COLOUR, which were not judged demonstrably differently by respondents likely to have such pronunciations in their own accents. This fits in with the results discussed in 3.5.8, which show that RED was generally considered to be a serious error. Similarly, it was suggested in 3.5.12 that the stigma attached to schwa epenthesis in FILM may also have affected Irish listeners, even though some of these may be expected to use similar realisations themselves. Clearly, certain realisations are so strongly stigmatised, even in speech communities where they are supposed to be salient, that it would be unwise to encourage their use by non-native learners of any variety of English.

Arguably, the stigma attached to realisations such as FILM and RED only applies to assessment of native English by native speakers. In these cases, native-speaker judges may well consider the same non-standard pronunciation feature to be more serious in other native speakers than in non-native speakers. In fact, research by Sebastian *et al.* (1978: 10–11, as discussed in Eisenstein 1983: 173) has revealed a tendency whereby accented speech which was considered to be Anglo-American was “rated lower” on the social scale than if the same speech was “accurately identified” as “Mexican American”. As Sebastian & Ryan (1985: 123) point out,

[t]hese findings accord with those reported by Fraser (1973) where black speakers were rated more negatively when misidentified as white than when accurately identified. The implication is that majority members hold a higher criterion for acceptable speech for Anglo Americans than for minority speakers.

Such a more lenient assessment of non-standard realisations in other ethnic groups could also extend to non-native speakers of English, such as the Dutch. This would suggest that Dutch learners would not have to worry about the stigma attached to FILM and RED. It should be remembered, however, that participants in the present experiment were explicitly told at the outset that they would be evaluating samples of Dutch English. In point of fact, many of them stated explicitly that they had taken this into consideration when judging the various tokens. If they still considered FILM and RED to be serious errors, this must be first and foremost be interpreted as their assessment of these tokens as they occur in Dutch speakers of English.

Clearly, some Dutch English realisations (including BOTH, THAT, WEATHER, CAR, and NEW) are judged less severely by speakers who may produce similar realisations themselves. In such cases, it may be less important for Dutch learners to insist on standard pronunciations of these tokens, at least when they interact with speakers of these accents. In other cases (such as FILM and RED), learners are well advised not to attempt realisations that may be too overtly marked as “non-standard”. As Swacker (1976: 17, quoted in Eisenstein 1983: 172) has claimed, “certain dialectal markers may be ... acceptable when coming from a native speaker, but quite offensive when spoken by a foreigner”. According to Eisenstein, these findings imply “caution in teaching the productive use of regionalisms to second language learners” (Eisenstein 1983: 172).

CHAPTER 5

CONCLUSIONS

5.1 Overview of the analysis of the two experiments

As was shown in 1.2.1, a significant number of studies have investigated the overall effects of accented L2 speech on different groups of L1 and L2 listener test subjects. A few such studies have also considered the different effects of various types of L2 pronunciation errors, such as those associated with segmental as opposed to supra-segmental features, phonemic as against sub-phonemic errors, or vowels versus consonants. On this basis, attempts have been made to prioritise certain errors in the pronunciation of, for instance, Swedish, German and Dutch learners of English (e.g. Johansson 1978, Dretzke 1985, Collins & Mees 2003b, Koster & Koet 1993). Such a hierarchy of error may be helpful in showing learners with these L1 backgrounds which pronunciation problems merit their greatest attention, and also provide new insight into the interface between phonetics, phonology and pedagogical approaches to second language acquisition (cf. Schwartz 2005). Recently, additional factors have also been included in considerations of error gravity, such as learner difficulty and proposed new prioritisations of learners' main target audiences – that is to say, whether non-native learners of English should be taught to communicate primarily with native speakers of this language, or with other non-native speakers. For instance, Jenkins (2000) has proposed a “Lingua Franca Core” of pronunciation problems that are significant for the type of English spoken primarily between non-natives. This “phonological core” excludes a number of phonemes (such as dental fricatives) that are considered a source of difficulty to most non-native speakers (Jenkins 2000: 138–139). This proposal, however, has been the subject of considerable controversy. In particular, the contentious notion that native-speaker norms are irrelevant to establishing priorities in foreign-language pronunciation teaching has encountered formidable resistance from native and non-native researchers alike (see Dziubalska-Kořaczyk & Przelacka 2005).

It is against this background that the present study has attempted to determine whether or not a pronunciation hierarchy of error can be established for Dutch learners of the two most commonly taught, and aspired to, accent models of English: Received Pronunciation and General American. As was also true of an earlier hierarchy of a similar type presented in Collins & Mees (2003b), this dissertation has emphatically taken the perceptions of native speakers of English as its starting point. Paradoxical though this may seem to some, such an approach is most likely to be in the interest of the non-native

learners. In spite of what Jenkins (2000) may claim, non-native learners of English are faced with the worldwide sociolinguistic dominance of its native speakers, and it would be irresponsible not to empower learners to deal with this inequality, especially if certain non-native accent features are subject to overt or covert stereotyping (see 3.7). In addition, Trudgill (2005a) has argued, based on McAllister (1997), that non-native speakers' comprehension of native-speaker English is dependent on their ability to pronounce its phoneme distinctions.¹ Perhaps most importantly, many learners, especially at an advanced level, do in fact model their speech on native-speaker English, or aspire to do so, and would object to being taught a non-native model instead (see Scheuer 2005: 126–127).

The focus of the present study has not been on representing the existing body of knowledge concerning native-speaker attitudes as discussed in pronunciation manuals. Instead, the primary aim has been to compare and contrast this knowledge with a multi-level statistical analysis of the actual evaluations of large numbers of native speakers of English with different cultural and linguistic backgrounds, and with dissimilar attitudes to the accent models in question. Given the abundance of varieties of English that exist, it is hard to conceive of English as a monolithic system, and equally difficult to view native-speaker evaluations of foreign accents as being dictated by a single and immutable set of norms that can easily be simplified to a common core. In view of this, considerable attention has been paid in this dissertation to discussing the significant differences in the detection and evaluation of pronunciation errors by dissimilar groups of native speakers.² Such differences were found to have an unmistakable impact on the hierarchies of error as constructed here for Dutch learners of Received Pronunciation and General American, in ways that could not be predicted simply from a comparison of the sound systems of these varieties of English (see 3.7 and 5.2.2).

Previous studies have occasionally referred to the linguistic background of native-speaker judges as a potential factor in the assessment of foreign pronunciation errors. However, the design of the core experiment of this dissertation, which involved the use of a web-based survey, has made it possible to elicit and compare such judgements, in a structured fashion, from a relatively large number of native-speaker respondents, including linguistically naive judges and speakers of varieties other than “standard” accent models such as RP and GA. To the present author's knowledge, this is the first time that such an experiment has been carried out. As noted above, this special emphasis on the sociolinguistic aspects of error detection and evaluation has helped to suggest

¹ For a recent discussion of the different factors which adversely affect non-native as opposed to native listening, see Cutler *et al.* (2004: 3674–3676).

² To describe variation in English, Gibbon (2005: 446–452) instead proposes Wittgenstein's (1953) well-known “Family Resemblance Model”. This would appear to do more justice to the complex interrelations between varieties of English than a “Common Core” (Jenkins 2000).

differences in approach and attitude between the various native-speaker groups (see 3.4.4 and 5.2.3). In particular, the tendency of North Americans to rate more severely than did other groups the smaller number of errors they apparently found “clearly detectable” may come as a surprise to some learners who might expect British judges to be stricter.

This sociolinguistic framework has also provided some insight into the effect on native-speaker judges of hearing foreign pronunciation errors that are actually similar to realisations with which they are already familiar. They may, for instance, recognise these as stereotyped foreign pronunciations or as the shibboleths of L1 regional or social accents other than their own. The results of the present study indicate that this has caused different groups of judges to evaluate such errors quite variably – sometimes more, sometimes less, severely – depending on the deviation in question. In those cases where native speakers perceive typically Dutch realisations as being similar, or identical, to L1 pronunciations heard in accents like their own, the effect has been referred to as “accent similarity”. Whereas it may be naively assumed that this will cause these groups of L1 judges to evaluate realisations of that type more leniently, the effect, while attested in a number of cases, has nevertheless been shown to be quite elusive (see 4.6 and 5.2.5). Clearly, one cannot simply assume that certain L2 pronunciations are acceptable only because they also occur in certain varieties of L1 speech. In other words, there is no support for Jenkins’s (2000: 27) peremptory statement that it is “no longer appropriate to regard ... variation from the L1 as automatically deviant” since “[m]uch of it comprises acceptable regional variation on a par with that which we find among L1 accents of English”. In fact, the results suggest that speakers of varieties other than RP and GA are, generally speaking, quite prepared to judge foreign-accented English by the norms of these two learner models, even where these are at variance with those of their own accents. This appears to be true even of respondents from Scotland, Ireland and Australia, where a certain degree of antipathy to accents such as RP is not uncommon (see 5.2.4).

The main findings of the core experiment show that intelligibility is not the sole criterion used by native speakers in deciding whether a particular pronunciation error is acceptable. Respondents’ emotive reactions to certain stigmatised realisations indicate that factors such as irritation or amusement also play a part in prioritising certain errors over others. The importance attached by different groups of judges to these additional factors in prioritising pronunciation errors is not sufficiently reflected in those studies which advocate “mere” intelligibility (e.g. Munro & Derwing 1995: 93) or which pointedly ignore native-speaker concerns (e.g. Jenkins 2000: 158). Nor do the findings lend support to the claim made by Albrechtsen *et al.* (1980: 395) that “one should not expect to establish a hierarchy of errors with respect to irritation”, since “[a]ll errors are equally irritating, provided they are ... violations of a target language norm”. Respondents’ comments indicate that, across different groups of native speakers, some errors are clearly and consistently more irritating than others.

The results of the core experiment also show that certain errors may be more or less salient depending on their position in the word. While phonologists and phoneticians are increasingly aware of the effects of syllable position on perceptual salience (Beckman 1999: 3, 20, Kingston 1985, 1990, Steriade 1993), this seems not to be generally reflected in any pronunciation manuals or studies of error gravity – except in the context of errors that are characteristically found in a particular phonological context (such as phonemes affected by final devoicing). Furthermore, the findings of the main experiment additionally suggest that contextual factors also account for the relative lack of priority given to certain pronunciation errors, for instance those of a suprasegmental nature.

Another practical objective of this dissertation has been to provide recommendations on a pronunciation error hierarchy for the benefit of Dutch learners of English. To this end, in addition to the core experiment with native speakers, another survey was set up in which teachers, lecturers and students involved in English-language education in the Netherlands were asked to evaluate a number of characteristically Dutch pronunciation errors. It was in fact on the basis of the assessments of the latter that certain errors were selected for inclusion in the core experiment. As a result, it was not only possible to make a balanced choice of those errors considered relevant by respondents working with English in Dutch secondary schools and universities, but a comparison was also thereby possible between their judgements and the assessments of native speakers of English. Consequently, the findings of this dissertation have direct bearing on the actual practice of English pronunciation teaching in the Netherlands, and could serve as a basis for the recommendations provided in 6.1 and 6.2 on how such training may be carried out with greater efficiency.

In spite of the similarities between the Dutch survey and the core experiment involving native speakers, there were also a number of structural differences which made it difficult to compare the two experiments reliably in all aspects (see 4.6 and 5.3). As a result, no clear evidence has been found to show that non-native speakers, and pedagogues in particular, are stricter judges than native speakers, as is sometimes claimed, or that these two groups unquestionably prioritise different types of error. Nevertheless, a comparison of the two experiments has helped to identify a number of complicating factors affecting how different groups of native and non-native judges evaluate errors. These include a tendency on the part of non-native respondents to attach too little importance to a number of both phonemic and sub-phonemic errors, and a disinclination to recognise how these errors are evaluated differently by native-speaker judges of the RP and GA versions respectively.

The Dutch survey contained a number of general questions on the status of pronunciation teaching as a component of English-language teaching in the Netherlands. A comparison of the answers provided by the secondary school-teachers and university lecturers (with regard to actual practices in their own institutions) and by Dutch students of English (with reference to their own experiences in secondary school) has shown that in secondary schools much less attention is given to all aspects of pronunciation training than is the case at

universities and colleges (see 2.2.1), in spite of the fact that pronunciation may be subject to evaluation at all educational levels. These findings raise serious concerns about Dutch secondary school pupils' lack of access to, and meta-linguistic awareness of, non-Dutch accent models of English, other than through haphazard linguistic encounters (see 2.2.1). A similarly alarming result was that a small minority of secondary school teachers do not actually require their pupils to speak English at all, either as a classroom activity or through any other procedure.

5.2 Summary of the main findings

5.2.1 Hierarchy of error: general principles

Although it is possible to establish a general hierarchy of error on the basis of native-speaker evaluations of Dutch pronunciation errors, it is more useful to provide separate hierarchies for the RP and GA version of the experiment. A great many errors were evaluated consistently differently by these two groups of judges, which indicates that these two pronunciation models should be treated as having clearly distinct priorities in error gravity. (The different hierarchies are presented in 5.2.2.) Despite the dissimilarities between the RP and GA forms, there were a number of general principles underlying the judges' assessments in all versions of the Native-speaker Experiment:

(1) Not all errors involving suprasegmental features are equally significant. While the errors involving word stress were considered to be among the most important, much less significance was accorded to the avoidance of weak and contracted forms, while intonation errors were rated among the least important. More research is required to establish whether this ranking reflects any intrinsic qualities of these features. It may also be the result of the context-free presentation of the errors, the particular selection of suprasegmental features, or a design fault of the experiment. Additionally, it could also imply that Dutch intonation in English is less distracting than that of certain other languages (such as, for example, Swedish). If so, it would suggest that caution is required in applying hierarchies of error developed for other languages to Dutch.

(2) Phonemic errors are not always more significant than sub-phonemic errors. It is true that there was a general tendency for phonemic errors to be ranked more highly than those of a realisational or distributional nature. Nevertheless, there were a number of clear counter-examples testifying to the significance of sub-phonemic errors. These were either errors which involved important acoustic cues or which evoked irritation or amusement. In addition, there were a number of classic phonemic errors (such as fortis/lenis neutralisation and the substitution of /ð/ by /d/) that were considered less significant when occurring in

high-frequency grammar words (see 3.5.6, 3.5.7 and 3.7). While it is true that some sub-phonemic features (such as aspiration and glottalisation) in fact help to support phoneme contrasts, nevertheless other sub-phonemic features (such as uvular realisations of /r/) do not perform this function in any way. This suggests that avoidance of phoneme conflation should not be the sole concern of pronunciation teaching.

(3) *Phoneme contrasts with a low functional load may still be significant.* If certain phoneme contrasts are attested in only a small number of minimal pairs, such as /ʊ ~ u:/, they are often assumed to be less important to intelligibility than oppositions that are found in a great many words (cf. Brown 1988, Wells 2005). This is not borne out by the significance attached to /ʊ ~ u:/ confusion in the RP form (see 3.5.22). Nor does it account for the considerable severity accorded to the substitution of /ð/ by /d/ in all versions of the experiment (see 3.5.6), despite Brown's (1988: 222) relatively low ranking of this phenomenon. Seemingly, functional load cannot be used to predict the significance of all phonemic errors.

(4) *Consonantal errors are not more significant than those involving vowels.* There is no evidence to support the notion that errors involving consonants are prioritised over those of a vocalic nature, as has been suggested in Johansson (1978: 97, 111), Koster & Koet (1993: 77) and Munro & Derwing (1995: 76). Some vowel errors were in fact ranked very highly (especially in the RP version of the experiment). For instance, all groups of native speakers concurred in assigning a high priority to /æ ~ e/ confusion (see 3.5.2).

(5) *Stigmatisation is a significant indicator of error acceptability.* Errors which had no appreciable consequences for intelligibility were nevertheless sometimes viewed as highly significant. Respondents' comments indicated that these errors were either stereotyped foreign pronunciations (such as the use of uvular-*r*) or stigmatised realisations associated with L1 varieties of English (such as schwa-insertion in *film*, or substitutions of /ð/ and /θ/ by dental stops). This indicates that foreign accents are not only judged on the basis of intelligibility, and that learners should also expect their accents to be evaluated by L1 standards for acceptability.

(6) *Some judges describe stereotypically "British" or "American" pronunciations as errors, but these are not regarded as very significant.* As almost all North Americans will pronounce words such as *secondary* with four syllables, some of them described the characteristically British trisyllabic pronunciation as an error, while some British, Irish and Antipodean respondents did the same for the North American realisation. Interestingly, none of the groups mentioned above attached a great deal of significance to this error. A similar case may be made for the British, Irish and Australian evaluations of what is considered to be a stereotypically American tendency towards yod-deletion

(in words such as *new*) – although some respondents were less tolerant of this realisation (see 3.5.16). Clearly, not all “Britishisms” and “Americanisms” should be accorded equally high priority.

(7) *The salience of errors may be affected by their position in the word or syllable.* This is suggested by the varied responses to TH-stopping in initial, medial or final position (see 3.5.5 and 3.5.6), and the different evaluations of /f ~ v/ confusion in initial or final position (see 3.5.3). However, they may well be other factors which account for these findings, such as an actor’s performance, word frequency or the absence or presence of a minimal pair.

5.2.2 Hierarchies of error for the RP and GA versions

For each of the two versions, estimates of respondents’ severity evaluations of all 32 tokens have been calculated by means of multi-level analysis. These have been ranked into a number of clusters, which comprise combinations of estimates that had not been evaluated significantly differently. While the resulting clusters in the RP version do not follow a clearly discernible pattern (see 3.2.4), the corresponding three clusters in the GA version consist of the following types of errors (see 3.2.5):

- an upper range of the **most serious** errors, with estimates exceeding 2.2 Likert scale points;
- an **intermediate** range with estimates between 2.2 and 0.5 scale points;
- a lower range of the **least serious** errors, with estimates below 0.4 scale points.

These ranges may be used as the basis for a more detailed hierarchy of error for GA, consisting of five main groups, as has been done in Table 5.1. Grouping bars indicate which clusters of errors are not statistically different from each other. A similar division was adopted, in the interest of comparison, for the RP version – although differences between RP clusters are much less clear-cut than in the GA form.

These five groups are intended as general indications of error severity, reflecting a number of significant differences between the RP and GA versions (as discussed in 3.2.6). They should not be interpreted to mean that all estimates in a particular group are statistically different from all other estimates in another group. For instance, the tokens representing phonemic consonant substitutions in high-frequency words such as *off* and *that* were not assessed significantly differently, in the RP version, from the token representing overlong /aɪ/ (for details, see 3.2.4).

Table 5.1. Suggested hierarchies of error for the RP and GA versions. The numbers used to define error clusters refer to severity estimates (expressed in Likert scale points), ranging from **most serious** (> 3.5) to **least serious** (< 0.4). Grouping bars denote error clusters that are not statistically different.

	Received Pronunciation	General American
(> 3.5)	Stress errors	Stress and stress-related errors Fortis/lenis neutralisation (f ~ v, t ~ d) Use of uvular- r
(2.2–3.5)	Stress-related errors Fortis/lenis neutralisation (f ~ v, t ~ d) Use of uvular- r Some substitutions of /θ, ð/ by /t, d/ Glottalisation of final /d/ Epenthetic [ə] in /lm/ /v ~ w/ confusion Confusion of /æ ~ e, ʌ ~ ɒ, ʊ ~ u:/ Unaspirated [t]	Most substitutions of /θ, ð/ by /t, d/ Glottalisation of final /d/ Epenthetic [ə] in /lm/ /v ~ w/ confusion /æ ~ e/ confusion Inappropriate post-vocalic r
(1.2–2.2)	Absence of weakening in <i>secondary</i> Absence of weak and contracted forms Inappropriate post-vocalic r Some substitutions of /θ, ð/ by /t, d/ Yod-deletion in <i>new</i> Degemination of /t#t/ Overlong /aɪ/	Weakening in <i>secondary</i> Phonemic consonant sub- stitutions in high-frequency words such as <i>off</i> and <i>that</i> Degemination of /t#t/
(0.4–1.2)	Phonemic consonant sub- stitutions in high-frequency words such as <i>off</i> and <i>that</i> Some intonational deviations	Absence of weak and contracted forms Unaspirated [t] Confusion of /ʌ ~ ɒ, ʊ ~ u:/ Overlong /aɪ/ Overdark pharyngealised [ɤ] Some intonational deviations
(< 0.4)	Some intonational deviations Overdark pharyngealised [ɤ]	Yod-insertion in <i>new</i> Some intonational deviations

The differences in evaluation of particular errors between the RP and GA versions allow for the following conclusions:

(1) *Americans and Canadians prioritise Dutch pronunciation errors structurally differently from other groups of L1 speakers of English.* In spite of the fact that there were a number of general patterns common to all judges (see 5.2.1), no fewer than 22 out of 32 tokens were judged significantly differently by Americans and Canadians. This included aspiration, which is commonly described as “essential” to the articulation of initial fortis plosives in American English (see Collins & Mees 1993: 14), but the absence of which was assessed much less seriously by North Americans (see 3.5.10). The fact that many of the latter failed even to detect unaspirated [t] makes one wonder if aspiration of initial fortis stops is a necessary acoustic cue for speakers of GA.

(2) *The different error assessments for RP and GA extend to suprasegmental phenomena.* Interestingly, the errors representing avoidance of weak and contracted forms were assessed significantly more leniently by North American respondents (see 3.5.18). Different attitudes to the various intonational errors were also attested (see 3.5.23).

(3) *The different priorities given to particular errors by judges of RP and GA cannot merely be predicted from the features that distinguish these varieties.* An example of this is North Americans’ assessment of /ʌ ~ ɒ/ and /ʊ ~ u:/ confusion, which was dramatically less severe than those of judges from Britain, Ireland and the Antipodes. This cannot be derived from a comparison of the phoneme inventories of GA or RP (see 3.7).

(4) *The stigma attached to particular pronunciations may be stronger, weaker or non-existent in either RP or GA.* This is, for instance, apparent from those errors that have the effect of consonant deletion or insertion. For example, r-retention where the prestige variety is non-rhotic is slightly less severe than r-deletion where the prestige variety is rhotic (see 3.5.13). Similarly, while L-vocalisation may be subject to some stigmatisation in the United States, assessments of overdark [ɫ] in the British Isles and the southern hemisphere suggest that L-vocalisation is not an issue with speakers of these varieties (see 3.5.20).

(5) *North Americans appear to attach a greater stigma to fortis/lenis neutralisation, and the replacement of dental fricatives by dental stops, than other groups of L1 speakers of English.* An explanation for this may be found in the association of these phenomena with the heavily stigmatised African American Vernacular English (see 3.5.1, 3.5.11, 4.4.1, 4.4.5, 4.4.10), and other accents subject to such stereotyping. The stigma appears to be much less apparent in high-frequency grammar words such as *that* and *off* (as was also found for RP).

If these results are compared with previous attempts to prioritise Dutch pronunciation errors, it appears that there is a large amount of consistency between the present study and the most detailed existing hierarchy of error formulated for Dutch learners of RP, as found in Collins & Mees (2003b: 290–291). Nevertheless, the present study shows that overlong /aɪ/ and dark [ɪ] are less significant than the authors suggest (although in the former case, the results may have been affected by the design of the experiment). While unaspirated [t] and confusion of /ʌ ~ ɒ, ʊ ~ u:/ are clearly very salient errors within the context of RP, Collins & Mees (1993: 124–130) appear to have overestimated the importance of these for GA. It should be noted, however, that the discrepancies between the findings of the present study and the various error hierarchies proposed by Collins & Mees (2003b, 1993) may well be the result of the employment of different methodologies (see 1.2.2).

The present results are more difficult to compare with other attempts to discuss the relative importance of Dutch pronunciation errors, such as Gussenhoven & Broeders (1997: 16–17) and Koster & Koet (1993), since these only mention some of the errors included in the survey, and do not provide an explicit hierarchy of error. However, the conclusion reached by Koster & Koet (1993: 90) that Dutch teachers are justified in paying little attention to suprasegmental features (such as intonation and the use of weak forms) has not been completely confirmed by the present experiments. Clearly, further research is required in this area. When the results of this dissertation are contrasted with Dretzke's (1985) investigation into an error hierarchy for German, this shows that the latter study attaches considerably less importance to fortis/lenis neutralisations, /æ ~ e/ confusion and incorrect stress. This inconsistency may well result from the use of very different methodologies, and the enrolment of dissimilar groups of respondents; Dretzke in fact drew his participants from secondary schools in just one city (Newcastle-upon-Tyne and its environs) situated in the north-east of England.

It is one of the objectives of this dissertation to compare and contrast the priorities for pronunciation teaching provided by Jenkins (2000) with the error hierarchies devised in the present study on the basis of native-speaker reactions to Dutch English. As has been pointed out before, Jenkins's suggestions were specifically made with a view to increasing intelligibility between non-native speakers of English rather than with any native-speaker interests in mind. In this respect, her aims may well be quite at variance with those of the majority of advanced Dutch learners, who no doubt also wish to use English in communication with native speakers. In the unlikely event that it is nevertheless decided to adopt Jenkins's suggestions wholesale in pronunciation teaching in the Netherlands, it would be useful to know whether the resulting Dutch-accented EIL could impair intelligible and efficient communication with native speakers.

It is one of the significant results of the Native-speaker Experiment that certain representative Dutch pronunciation errors are prioritised differently by different groups of native speakers. In some cases, this is linked to dissimilar levels of stigmatisation accorded to certain pronunciation features by, for

instance, North Americans, as opposed to the reactions of native speakers of other varieties of English. As Jenkins does not purport to be concerned with native-speaker norms or the stigmatisation of specific pronunciations, these differences in native-speaker evaluation have not been factored into her recommendations. This means that some of her suggestions (such as substitutions of dental fricatives by stops) will adversely affect Dutch learners' communication with North Americans in particular, whilst others (such as aspiration of initial /t/) will have an especially significant effect on judges from the British Isles and the Antipodes. In fact, the findings of this dissertation suggest that aspiration is not in any way a high priority for speakers of GA.³ Similarly, while most American and Canadians are unlikely to object to a consistently realised post-vocalic [ɹ] as recommended by Jenkins (2000: 139), this is not quite true of all other groups of respondents. In addition, certain sociolinguistically sensitive pronunciations produced by Dutch learners (such as uvular-*r*, and the use of epenthetic [ə] to break up clusters) are not even mentioned in the Jenkins's Core, even though these were ranked among the serious errors by all groups of native speakers. Furthermore, the proposals made by Jenkins appear to ignore the complex reactions found in some groups of native speakers to the non-native use of pronunciations similar to those heard in their own speech community (see 5.2.4 and 5.2.5).

Apart from the fact that Jenkins does not warn learners against using realisations that may be sociolinguistically marked, or which may evoke irritation in particular groups of native speakers, she also fails to emphasise sufficiently the importance of crucial phenomena such as word stress and the maintenance of certain phonemic distinctions. The results of the present study show that native speakers do not merely consider word stress to be "reasonably important", as Jenkins (2000: 150) claims, but view its incorrect use as one of the most significant errors. Similarly, while Jenkins (2000: 159) correctly but somewhat vaguely suggests that learners should avoid substituting English consonants by "certain approximations ... where there is a risk that they will be heard as a different consonant sound from that intended", she does not extend the same admonition to most vocalic phoneme contrasts, where "L2 regional qualities [are] permissible if consistent". Since it is a characteristic of the "regional" English spoken in the Netherlands and Dutch-speaking Belgium that phoneme contrasts such as /æ ~ e, ʌ ~ ɒ, ʊ ~ u:/ are merged or confused, this would imply that Jenkins's recommendations are totally inconsistent with the significance attached, by native speakers, to the preservation of these distinctions, especially as far as the /æ ~ e/ contrast is concerned.

³ In any case, aspiration is certainly not "particularly important" to non-proficient non-native speakers, as Jenkins (2000: 140) incorrectly claims, as this surely holds only true for those whose L1s employ aspiration as an acoustic cue. The same point may be made about Jenkins's (2000: 140) suggestion to incorporate the "differential effects of fortis and lenis consonants on the length of a preceding vowel sound" in the *Lingua Franca Core*. This is unlikely to be important to non-native speakers whose languages do not employ such effects.

Some of the errors discussed in the present study do not feature, or only obliquely, in Jenkins's Core, and cannot therefore be usefully compared. However, the results of the present experiment with regard to weak forms and intonation appear to be in accordance with the low priority given to these features by Jenkins (2000: 146–156). Nevertheless, it should be pointed that the importance of weak forms was in fact assessed differently by various groups of native speakers, and that the intonation results were not totally conclusive. In other words, the findings of this study do not lend ample support either for or against Jenkins's recommendations in these matters. However, it may be noted that in at least one respect, the suggestions made by Jenkins appear to be clearly consistent with the native-speaker evaluations. The results suggest that Jenkins (2000: 138–139) may well be justified in assigning a low priority to “the production of dark [ɫ]”, especially with regard to British English (as was also noted by Wells (2005: 105) in his review of Jenkins's recommendations).

Not only are there striking discrepancies between the proposals made by Jenkins for the purpose of non-native communication and the native-speaker reactions discussed in this dissertation, but her recommendations, if followed up, could adversely affect linguistic interactions between Dutch learners of English and native speakers (even though this would be of little concern to Jenkins, who does not prioritise communication with native speakers). If intelligible and efficient communication between such groups is still considered to be an important goal of English teaching, it would be inadvisable to adopt the *Lingua Franca Core* in pronunciation training in the Netherlands.

5.2.3 Detection as a factor in error assessment

An important factor underlying differences between groups of judges is the influence of the separate effects of the error detection success rate (“hit rate”) and of the severity assessment of those errors actually detected by the respondents (“adjusted severity”). Such effects are particularly different for male as opposed to female respondents, younger as against older judges, and North American participants versus those from the British Isles and the southern hemisphere (for details, see 3.4.4). It is, however, only if both factors are taken into consideration that justice can be done to the actual significance of an error. That is to say, an error can only be ranked in a hierarchy if this is based on the combined effects of detection and assessment. This is the “composite severity estimate” which has informed the hierarchies of error in 5.2.2. Nevertheless, an analysis of the individual effects of “hit rate” and “adjusted severity” may lead to the following conclusions:

(1) *Respondents' success in detecting an error is not necessarily linked to their assessment of its severity.* Some errors were considered to be important, but were not widely detected, whereas others were reported by a great many judges, but were not described as in any way serious. While the former suggests that there are groups of native speakers with “idealised” pronunciation norms for errors that are hardly detected by the vast majority of respondents with similar

accents, the latter suggests that some groups of judges will not hesitate to report certain errors whilst simultaneously denying their significance. Although respondents' behaviour may have been modified by the instruction given at the outset of the Native-speaker Experiment only to consider "clearly detectable" errors, these tendencies may nevertheless also reflect structurally varying attitudes, or dissimilar strategies, to error detection and assessment in different groups of respondents. Such attitudes or strategies are clearly worth further investigation.

(2) *British, Irish and Antipodean judges report more errors than North Americans, but diagnose fewer of those as serious.* This may point to an overall attitude to Dutch English, or possibly to non-standard or non-native accents in general, which may be defined as "noticeable but not serious". On the other hand, it may suggest that such judges (with the possible exception of the Irish respondents) are less inclined to admit to intolerance of pronunciation errors than North Americans. Be that as it may, this result is inconsistent with the Dutch perception that British and Irish native speakers are stricter judges of Dutch English pronunciation than Americans and Canadians.

(3) *North American respondents detect fewer errors, but evaluate those detected as more serious.* This tendency may well reflect a general perception of Dutch English, or potentially of regional or foreign accents in general, as "serious only where noticeable". This indicates that Americans and Canadians are more prepared to volunteer negative evaluations of accented speech than their European and Antipodean counterparts. It would be mere speculation to ascribe this simply to greater ethnocentricity. In any event, it also means that North Americans tend not to detect or report all Dutch errors as frequently as some other groups do.

(4) *Increased error detection rates and/or higher severity assessments account for some of the differences in error types assessed less leniently by Americans and Canadians.* These include a number of characteristically Dutch errors which may have the effect of (1) neutralising the fortis/lenis contrast, of (2) substitution of dental fricatives, or (3) of consonant weakening or deletion (as in the codas of *tell* and *car*). It is likely that these were detected more frequently and/or evaluated more strictly in North America *not* because they lead to unintelligibility, but because of associations with stigmatised language varieties (see 4.4.1, 4.4.5, 4.4.10, 4.4.12, 4.4.19).

(5) *As a rule, women judge the errors they detect more severely than do men.* While female respondents' adjusted severity estimates were consistently higher than those of male respondents, the latter tended to have significantly higher detection scores. This result parallels the behaviour of North American respondents as opposed to that of other groups, and appears to be consistent with what Labov (2001: 266) has termed the "general linguistic conformity of women",

causing them to judge whatever they consider to be deviant from the norm more stringently than do men.

(6) *Younger respondents tend to detect more errors, and assess them more severely, than do older respondents.* This may be ascribed to the older judges' "aging auditory system" (Sommers 2005: 469) or to their greater experience with, and tolerance of, "language variations" (Ryan 1983: 154).

The following considerations should be noted. Firstly, a greater acceptance of foreign-accented English may also be affected by a number of additional factors (see 3.7). Apart from a greater awareness of L1 linguistic variation, a tendency towards romantic or practical appraisals of foreign speech, or an indebtedness to non-native speakers for speaking a foreign language (cf. Nickel 1972: 19–20, Johansson 1978: 119), such factors may also include covert motivations that are exclusionist rather than integrative (see Prator 1968: 25, Leather & James 1996: 271, Scheuer 2005: 112). Some native speakers may even be motivated by a desire to excuse or vindicate their own monolingualism. Seen in this light, negative evaluations of non-native accents could also be construed as being indicative of a more matter-of-fact attitude to foreign accents, inspired by integrative views of immigration.

Secondly, it must be remembered that, while some groups may be stricter judges of the errors they reported, this does not mean that they will necessarily evaluate pronunciation errors more negatively in, for instance, a classroom situation. If their detection scores are also significantly lower, this may well compensate for their less lenient assessments.

5.2.4 Error assessment in the different accent groups

The differences in error assessment between accent groups can mostly be accounted for by variation between the RP and GA versions of the experiment. The most important of these were discussed in 5.2.2. In addition, there were a number of interesting results for each of the major accent groups, which are discussed below.

British English

There were very few differences between British respondents who described themselves as speakers of RP, or Standard Southern British English, and those who did not – apart from a tendency towards slightly varying assessments of uvular-*r*, glottalisation of lenis consonants, fortis/lenis neutralisation in *off*, and avoidance of vowel gradation in *that*. This indicates a willingness on the part of non-RP speakers to judge Dutch English pronunciation errors by the ex-normative standards of RP (insofar as these are different from regional accents). Although judges from Scotland tended to be slightly more lenient (and also described themselves as such), by and large they appeared to evaluate errors

according to the same, or sometimes even stricter, norms. Interestingly, this included the Dutch conflation of /ʊ ~ u:/, despite the fact that this is also a feature of educated Standard Scottish English (see 3.5.22, 4.6). Such awareness of RP norms is perhaps remarkable, given the negative feelings towards this accent in Scotland (McClure 1994: 80). This relatively low regard, however, may explain Scottish respondents' greater leniency in this experiment, if it is assumed that such respondents would be generally more inclined to be tolerant of deviations from a target accent with which they do not identify.

Irish English (Northern and Southern)

As with British speakers of regional accents, Irish respondents appeared to be prepared to evaluate Dutch pronunciation errors by RP standards rather than their own, or at least to incorporate their awareness of linguistic variation in their judgements. Amongst other things, this is evident from the fact that, even though post-vocalic *r*-insertion is clearly not an error in Irish English, Irish respondents judged the relevant token no differently from the non-rhotic speakers of RP. Two errors resulting in fortis/lenis neutralisation (possibly associated with stigmatised foreign accents) were even assessed more strictly by Irish judges than by some of the other groups in the RP version. In fact, what appears to distinguish Irish participants more than anything else from their British and Antipodean counterparts is their greater willingness to criticise such pronunciation errors. Even realisations stereotypically associated with Ireland, such as schwa epenthesis in *film*, were sometimes described in very negative terms (see 3.5.12). This is not so to say that Irish judges evaluated this error as strictly as the other respondents in the RP form, but the relatively few respondents who detected it did appear to be aware of its stigmatisation.

Australian, New Zealand and South African English

There was an overall tendency for judges from Australia, New Zealand and South Africa to assess the errors slightly more leniently than all other groups of native speakers combined. As in the case of the Scottish judges, this may be related to a reduced accent loyalty for the British prestige variety employed in this experiment. It should be noted that this trend towards greater leniency is not very pronounced. For example, Antipodean judges' less severe evaluation of /æ ~ e/ confusion is only significant when compared with all other groups put together (see 3.5.2). There were few other errors that were assessed significantly differently from any of the other groups in the RP version.

American English

There were no significant differences between any of the major accent groups in the GA version. Nevertheless, there were a few possible indications that respondents who described themselves as speaking a variety other than GA evaluated certain fortis/lenis neutralisations (and substitutions of /ð/ by /d/) more negatively (see 3.4.2). Whilst in some cases this may be ascribed to greater linguistic insecurity, it could also point to an increased awareness of the social consequences of using stigmatised speech. It is also interesting to note that judges who labelled themselves as being from the American East Coast tended to be stricter, and were also more inclined to describe themselves as such. It is possible that these respondents identify more strongly with the standard language than do many other groups.

Canadian English

Given the many similarities between mainstream Canadian English and GA, it is hardly surprising that there were no significant differences between respondents from Canada and the United States. There was, however, a possible tendency for Canadians to evaluate overlong /aɪ/ in *ice* somewhat less strictly than did their US counterparts (see 3.5.9). This may be related to the phenomenon of Canadian Raising (Wells 1982: 494–495), as a result of which Canadian respondents may be less inclined to detect or reject any realisations of /aɪ/ that do not conform to GA norms. Interestingly, while no Canadian provided a negative comment on any of the errors, their assessments were no less strict than those of US respondents.

The findings from these groups result in the following conclusions:

(1) *It is not only speakers of the supra-regional standard variety (such as RP or GA) who are ready to judge foreign pronunciation errors by the norms of this particular variety.* For instance, a readiness to judge a Dutch accent by RP standards was attested in regions as different as Scotland, Ireland and the Antipodes. This suggests that Wells's (1982: 279) observation that "[e]veryone in Britain has a mental image of RP" also takes in other English-speaking countries.

(2) *There is a general tendency for speakers of some regional varieties to be slightly more lenient than speakers of supra-regional standard varieties.* This phenomenon could be related to the degree to which respondents feel loyalty for the accent model used by the learner.

(3) *There may be a tendency for speakers of some regional varieties to be stricter about particular stigmatised pronunciation errors than are speakers of supra-regional standard varieties.* This effect could be motivated by diverse factors, including linguistic insecurity, or even an unwillingness to accept that foreign learners may wish to speak a local variety of English, but also a greater awareness of adverse reactions to regionally flavoured or stigmatised speech.

5.2.5 Accent similarity

The following conclusions may be drawn:

(1) *A large number of characteristically Dutch pronunciation errors in English are similar or identical to realisations heard in regional speech from the British Isles, North America, Australia, New Zealand and South Africa.* As the overviews in 4.2 and 4.4 show, at least 20 mostly segmental errors included in the native-speaker survey correspond to realisations heard in a minority of speakers in different accent groups. Four of these (raising of /æ/ to [ɛ], r-deletion or r-insertion, unaspirated [t] and conflation of /ʊ ~ u:/) are in fact associated with a majority of speakers in one or more accent groups.

(2) *If pronunciation errors are similar, or identical, to realisations heard in respondents' own accent groups, the latter generally tend to evaluate these errors more leniently.* This effect was not only attested for the 20 relevant errors combined, but was also found for three /ð/ substitutions, and for the insertion/deletion of /r/ and /j/ (in words such as *car* and *new* respectively). This implies, for instance, that non-rhotic speakers of North American English are very likely to judge r-deletion in GA less leniently than the vast majority of rhotic speakers; the same may be expected of rhotic speakers' assessments of r-insertion in RP. While this may not come as a surprise, it is a useful reminder that some regional speakers have a different hierarchy of error from speakers of the "standard" or prestige variety.

(3) *Some such pronunciation errors are evaluated no differently than they are by other groups of judges.* There was no observable effect of "accent similarity" on as many as 14 out of 20 tokens, including some realisations that are very common in some accent groups. In at least three cases (/ʌ ~ ɒ/ conflation, uvular-r and schwa epenthesis), the absence of any such effect cannot easily be attributed to large sampling errors. Errors such as these are likely to be strongly stigmatised, even in speech communities where they are supposed to be salient accent features.

(4) *At least one such error was evaluated more severely by judges who are likely to be produce similar realisations themselves.* The conflation of /ʊ ~ u:/ was assessed more strictly by respondents from Scotland and Northern Ireland. This may be due to a number of reasons (see 4.6), but it may still suggest that "accent similarity" in itself is not the sole factor in accounting for judges' evaluations of pronunciations similar or identical to their own.

(5) *Dutch and other foreign learners of English should not be encouraged to imitate certain regional or local accent features merely because these are very similar to their own characteristic non-native realisations of English.* It would be unwise to do so without any awareness of the reactions such realisations are likely to engender in speakers of such accents (or in other native speakers). An exception should be made for those errors that have been shown to be well-received, or hardly detected, by the group of L1 speakers the foreign learner is attempting to interact with. For example, the retention of post-vocalic /r/ is clearly unlikely to present any problems of intelligibility or acceptability in interactions with speakers of rhotic varieties of British English, provided, of course, that the particular realisation of /r/ is not heavily stigmatised, and that the learner pronounces it consistently.

5.2.6 Comparison with the Dutch Experiment

Insofar as the two experiments can be compared at all (see 5.3), the following conclusions may be drawn about any differences between the native speakers of English, on the one hand, and those engaged in English language teaching in the Netherlands on the other (i.e. secondary school teachers of English, students of English, and lecturers in English departments of universities and colleges):

(1) *There was no clear evidence to show that native speakers consistently evaluated Dutch pronunciation errors more or less severely than did the Dutch teachers, students or lecturers.* Especially students' evaluations appeared to be quite consistent with native-speaker judgements – the secondary school teachers' assessments only slightly less so (see 3.6).

(2) *Dutch secondary school teachers may well have a slight tendency to underestimate certain sub-phonemic errors.* These include aspiration (if they teach RP) and glottalisation of lenis stops (especially if their model is GA). It may be noted that both aspiration and glottalisation are usually regarded as important acoustic cues for the perception of the fortis-lenis contrast by native speakers. Other possibly underrated pronunciation errors include degemination and a number of stigmatised realisations in GA (see 3.6).

(3) *There was a weak tendency for lecturers at Dutch colleges and universities to overestimate the importance of a few errors, and to underrate a number of others.* Potentially underestimated errors include the glottal replacement of lenis stops, and, from the point of view of the North American judges, r-deletion and incorrect phrasal stressing. From the perspective of the British, Irish and Antipodean judges, there was also a similar tendency to overestimate the importance of a number of other errors (see 3.6).

(4) *Dutch students of English may be slightly inclined to overemphasise the significance of a few pronunciation errors.* These included /ð ~ d/ substitution in

a high-frequency item such as *that*, and, if their model is RP, overlong /aɪ/ in *ice*. Some students appeared to underestimate the strength of North American objections to uvular-**r** (see 3.6).

(5) *In general, Dutch respondents' evaluations appeared to be unaffected by the different attitudes to certain pronunciation errors attested for RP and GA judges in the Native-speaker Experiment.* This is suggested by the Dutch participants' closer adherence to GA norms in some cases, but to those of RP in others. This implies that teachers, students and lecturers engaged in English language-teaching in the Netherlands will find it helpful to distinguish between RP and GA when it comes to prioritising pronunciation errors. They should also consider to what extent their evaluations of certain errors correspond with those of the L1 speakers whose accents they may be using as a model.

Given the differences between the two experiments, the above results should be regarded as tentative. In addition, they reflect the theoretical views of those teachers, lecturers and students who were actually prepared to participate in this survey on pronunciation. In view of the emerging evidence that little attention is given to pronunciation training in secondary schools (see 2.2.1 and 2.2.2), it is unclear how firmly teachers' theoretical pronunciation priorities are actually anchored in their teaching. Participants' responses to questions on the status of English pronunciation training revealed the following:

(6) *English pronunciation training is given much less priority in Dutch secondary schools than at universities and colleges, even though pupils' accents may be evaluated in all years.* This was true of all aspects of such training, including any kind of contrastive analysis of the sounds of English and Dutch, any discussion of, or reference to, pronunciation models such as RP or GA, and any actual exposure to the instructor's English (whether strongly accented or otherwise).

(7) *A small minority of secondary school teachers do not make their pupils speak English as a classroom exercise.* This may be the result of overburdened programmes, unrealistic class sizes or the desire not to tax their pupils' abilities unnecessarily. Nevertheless, this is likely to have serious consequences for these pupils' fluency skills in English.

(8) *A small minority of secondary school teachers of English appear to have "mostly Dutch" accents in the language they teach.* This would only be appropriate if a deliberate policy of teaching non-native English were to be adopted in the Netherlands.

5.3 Limitations of the present study

5.3.1 Introduction

There were a number of problematical aspects to the present study which suggest that a certain degree of caution is needed in interpreting its results. Some of these are inherent in research of this nature, or in the design of the experiment, whereas others represent possible flaws which only emerged as a result of post-hoc analysis. These different types of problems will be discussed in the sections below: while 5.3.2 to 5.3.4 relate to the core experiment involving native speakers, 5.3.5 will focus on the dissimilarities between this and the Dutch Experiment.

5.3.2 Presentation of the stimuli

The errors in the core experiment were presented to the native-speaker judges as a single deviation in an otherwise correct carrier sentence. This had the clear advantage of allowing linguistically native respondents to detect and assess foreign pronunciation errors without any prompting or verbal descriptions. However, it is clear that there may also be a number of disadvantages associated with this procedure.

Firstly, it may have caused these judges to underestimate the severity of these errors outside these isolated contexts. After all, strong foreign accents are characterised by a “layering” of errors, deviations and inconsistencies rather than one single unexpected feature (see Abbott 1991, Collins 1979b, Prator 1968: 19). Secondly, the degree of irritation produced by some errors (especially those of a suprasegmental nature) is likely to be more apparent upon repetition. It is also the interaction between different phenomena that is likely to obscure the severity of individual errors (as in the case of weak forms, stress and vowel gradation, see 3.5.18). The single deviations from what are otherwise perfectly acceptable RP or GA versions of the carrier sentences are actually more likely to remind native-speaker respondents of L1 regional variation – a consideration which may affect their attitude to these errors. Thirdly, the fact that respondents were also presented with a “clickable” version of the carrier sentence may have obscured any intelligibility problems which could have been caused by the error without a written context (as in the case of overlong /aɪ/, see 3.5.9). Fourthly, another factor that may have predisposed respondents to be more lenient is the presentation of suprasegmental errors without any disambiguating context to rule out any possible alternative interpretations of these carrier sentences as non-deviant; this was a particularly problematical in the case of the intonation tokens (see 3.5.23). Finally, it is not impossible that the salience of certain errors (within an otherwise correct carrier sentence) was reduced if respondents’ audio players played the stimuli imperfectly – an almost inevitable problem with experiments of this nature.

Such limitations, which are inherent in the design of the experiment, are likely to have decreased respondents’ ability and willingness to detect and assess certain or all errors more severely. This should be taken into consideration in

any analysis of these results, in particular when it comes to the tokens most affected by this. For instance, it is quite possible that with a different experimental design, the suprasegmental errors would have been evaluated much more strictly – although they may still have proved difficult to detect and assess for linguistically naive native-speaker judges. Similar reservations should apply to intonation. In retrospect, it may be stated that it is a design fault of the core experiment that the intonation component was not presented differently from the segmental and other suprasegmental errors. Since intonational deviations affect the entire utterance, it may be argued that, in each instance, an intonationally non-deviant version of the carrier sentence should also have been provided.

In addition, post-hoc analysis revealed that, in at least three carrier sentences, there were slight variations in prosody between the RP and GA versions of the experiment (see 3.5.5, 3.5.10 and 3.5.18). Furthermore, there was at least one case of unintentional variation in segmental realisation between the two versions (see 3.5.22). If such cases are ascribed to differences in performance between the two actors, this would suggest that some inter-version differences are not solely due to dissimilarities in error detection and assessment between the North American and the other L1 judges. This is an instance where the employment of different actors to produce different guises may have affected the outcome of the experiment.

5.3.3 Selection of participants

One limitation of the core experiment is its sampling bias. As in all online experiments (including the Dutch survey), only volunteers able and willing to complete electronic questionnaires have taken part in the survey. In addition, a large number of these respondents were drawn from the academic community and their relations – in other words, computer-literate, highly educated people who were interested in the experiment were much more likely to participate. This is bound to have consequences for the representativeness of the survey and respondents' degree of leniency. If, as one participant (Subject 902) pointed out, “[e]ducated English speakers understand that non-English speakers have difficulty with ‘th’”, this could be taken to imply that other groups of native speakers are inclined to be less tolerant of certain realisations.

Such leniency could also be the result of the fact that many groups of judges were asked to evaluate accent models (i.e. GA and RP) other than their own. If a separate accent guise had been made available for each relevant accent group, this could also have made it more attractive for certain categories of respondents (such as Australians) to take part in the experiment. Even though the number of guises was limited only because of the practical problems involved in finding suitable actors, the resulting effect of self-selection sampling may have meant that a significant proportion of native speakers, including many of those speaking emerging varieties of English from the Indian subcontinent and West Africa, did not participate. Consequently, their views on Dutch English pronunciation errors are not to be found in this survey. Similarly, since no respondents described their own linguistic background as African American

Vernacular English, it could not be established if such speakers respond differently to the Dutch accent features that other North American respondents appear to associate with AAVE. Within the framework of “accent similarity”, it would have been interesting to see whether or not self-identified speakers of AAVE also object to stigmatised features of their own accent when used by foreign learners, as was attested in, for instance, some Irish respondents’ reactions to schwa-epenthesis in the coda of *film* (see 3.5.12).⁴

A related problem in the core experiment was the use of respondents’ own accent self-identifications. This system was preferred because of the well-known mismatch between labels used by accent researchers and the public at large, yet some respondents’ self-identifications are clearly easier to interpret than others. In those cases where the labels were ambiguous, described “hybrid accents”, or referred to insufficiently well-documented varieties of English, respondents’ submissions were excluded from further consideration. This will have resulted in a certain sampling bias, the effects of which are difficult to estimate. In addition, the use of self-identifications may also have resulted in respondents describing their accents inaccurately, unclearly, or as more consistent with the prestige variety. While this effect is perhaps inevitable with experiments of this nature, it should be borne in mind in any analysis of the effects of categorising respondents into different accent groups.

5.3.4 Error detection

Since some respondents did not locate the phoneme or phonemes affected by the intended error, but selected adjacent phonemes in the same word or phrase (sometimes as a result of spelling confusion), or described the intended errors in their comments, it was decided that these methods of error detection would also be considered acceptable, provided no multiple errors had been evaluated simultaneously. Similarly, some participants appeared to have difficulty identifying suprasegmental errors, especially in terms of assigning these to either the “word stress” or “intonation” error categories. In view of this, both labels were accepted as description of these errors, as well as any relevant attempts to locate these in a particular phoneme or phrase, and any other unambiguous references as provided in respondents’ comments (see 2.5.2 for details). Although this was done in order to do as much justice as possible to participants’ submissions, it is not unthinkable that in a few cases, this may have led to an overgenerous treatment of error detection. Arguably, this is a disadvantage of the method employed in allowing respondents to identify errors in the core experiment.

⁴ If speakers of AAVE do indeed object to such features, this would go against the claim made by Lippi-Green (1997: 179) that “black concerns” about this variety of English “focus almost exclusively on grammatical issues”, whereas “whites seem to be most comfortable voicing overt criticism about phonological matters and sometimes about grammar”.

Certain respondents reported the intended errors in their comments whilst simultaneously refusing to categorise them as “clearly detectable”. This somewhat ambiguous category has been treated as distinct from those cases where respondents decided not to report anything, or volunteer any comments, even though they may well have been aware of any deviation or markedness. This implies that the question “Does this sentence contain a clearly detectable error?” may have been interpreted differently by different respondents. Whilst it is unclear whether this has affected the error detection scores, it serves as a reminder that reporting a detected error is not the same as merely being aware of it.

5.3.5 The Dutch Experiment

For the reasons explained in 3.6, there were considerable differences between the experiment aimed at native speakers and the earlier survey directed at Dutch respondents. This makes it difficult to compare and contrast the error evaluations of these two groups of respondents with any degree of confidence. For instance, there was a fundamental difference in tasks: while the Dutch participants were required only to assess somewhat technical descriptions of errors using one-word examples, the native speakers were not only supposed to evaluate errors in audio recordings of full carrier sentences, but also to detect the errors concerned. Although there was considerable overlap, the two experiments also employed a number of different errors and a different number of distractors (see 2.4.3). While the native speakers were invited to take part in either an RP or a GA version of the survey, no reference was made to these models in the Dutch version. Whereas the L1 judges identified their own accents, estimated their own leniency and stated their precise age, the Dutch participants supplied information about their educational background by selecting either the teacher, student or lecturer version of the experiment.

In view of these dissimilarities, it is, of course, only the adjusted severity estimates of the two main groups that can be validly compared, and then only for a limited number of errors. It should be remembered that the presentation of the stimuli in the core experiment may serve to increase respondents’ leniency, while this is unlikely to be a factor in the Dutch version. If, therefore, there are no striking differences in adjusted severity estimates between the native-speaker and Dutch surveys, this fact does not in any way suggest that both groups are equally lenient. Not only were the effects of error detection not factored in, but the Likert scales were also not anchored across the two experiments. In fact, what appears to be a consistently different use of the Likert scale implies differences in evaluation between these two groups, or at least between the different surveys in which they took part. Needless to say, such structural differences can only be proved to exist if a similar group of Dutch respondents also took part in the native-speaker survey – a possible recommendation for future investigations to be carried out in this area of research.

Finally, it should be mentioned that Dutch participants’ answers to the general questions about the status of pronunciation teaching in the Netherlands may well be influenced by the various groups’ dissimilar backgrounds, and

different motivations for taking part in the experiment. For instance, while some students of English may well have wished to criticise their own secondary school education in English, this is unlikely to have motivated the other groups of respondents to the same degree. In fact, all respondents may have been more than usually interested in pronunciation training in order for them to be motivated to take part in the survey. If this sampling bias has affected the results, this would suggest that pronunciation training is considered even *less* important by other teachers and students in the Netherlands. In view of the importance attached to a non-distracting and socially acceptable accent by native speakers, this is all the more reason to make pronunciation teaching a structural component of the curriculum at any level (see 6.2).

CHAPTER 6

RECOMMENDATIONS

6.1 Recommendations for future research

The present study has shown that Dutch pronunciation errors in L2 English are reported and assessed differently by different groups of native speakers. This may reflect fundamentally different attitudes to error detection and assessment, to foreign or regional speech, or to the pronunciation model used by learners. In addition, some errors may be stigmatised in certain groups but not in others. Where present, stigmatisation is likely to have a stronger effect on error evaluation than any similarities between Dutch pronunciation errors and authentic realisations produced by native speakers with accents like their own. These findings stress the importance of a sociolinguistic context for research into native-speaker evaluations of L2 speech, which has hitherto been largely ignored in pedagogical descriptions of error gravity.

The results above suggest interesting new avenues for research, and a range of replications of the current experiments in different formats. It is important to bear in mind that the findings of the present study are based on the willingness of a large and diverse group of native speakers to judge Dutch English pronunciation by the standards of Received Pronunciation and General American (see 5.2.4). For instance, if the core experiment were to be replicated with additional guises, such as Irish English, Australian English or African American Vernacular English, this might not only encourage considerably more speakers of such varieties of English to participate, but also provide more insight into the way judges from these groups assess Dutch pronunciation errors by the standards of their *own* accents (see also 5.3.3). This should be particularly interesting from the point of view of accent similarity, especially since the effect of this on error evaluation has turned out to be somewhat elusive in the present study (see 5.1).

Replications of the present study with additional guises could also be useful in establishing hierarchies of error for varieties of English other than RP and GA. Since it has been found that the hierarchies for GA and RP cannot be predicted from the features that distinguish these accents (see 5.2.2), it might be interesting to establish if this applies to other varieties as well. These hierarchies would also be relevant to those learners of English who wish to integrate into communities where RP and GA are not commonly used (for example, in Australia or Ireland). Such learners could benefit from an awareness of the priorities given to particular pronunciation problems by the native speakers

whose accents they may be attempting to imitate.¹ This information can be usefully included in any textbooks aiming to teach learners the pronunciation of these varieties. Since no detailed pedagogical descriptions of accent models other than RP and GA are available on the Dutch market (and seemingly nowhere else, to the author's knowledge), this appears to be a lacuna waiting to be filled.²

In view of the dissimilarities between the core survey and its Dutch counterpart, which has made it difficult to compare these reliably in all aspects, it might be rewarding to replicate the core survey with L2 speakers of English resident in the Netherlands and the Dutch-speaking part of Belgium. While one would expect Dutch respondents' error evaluations to be affected by lower detection scores, one might speculate that their greater familiarity with characteristically Dutch English mistakes will actually lead to increased severity (see also Koster & Koet 1993: 90). The likelihood of these effects could be established if the core experiment, or a similar version, were to be replicated with Dutch-speaking participants.

It would also be interesting to discover whether *other* groups of non-native speakers also evaluate certain Dutch pronunciation in English similarly to the L1 speakers of English. It has been amply demonstrated (Major *et al.* 2002, 2005) that non-native speakers of English find it harder to understand L2 English than native speakers. As Trudgill (2005a: 219) has argued, this is because non-native speakers "have greater difficulty in coping with the absence of phonological contrasts than natives". It would be helpful to see this notion confirmed in a replication of the present study, especially since it seems to be routinely ignored by proponents of English as an International Language (see also 3.7).

As was pointed out in 5.3.2, the method employed for the presentation of the audio stimuli entailed a number of limitations, as a result of which participants in the core experiment may have been led to under-assess the severity of a number of errors. It is to be recommended that, in any attempts to replicate the survey, the advantages and disadvantages of this method are carefully weighed up against any alternative approaches. In any replications, but in particular those involving non-native speakers, issues such as the absence or presence of a context, written or otherwise, and the "layering" of errors (Abbott 1991, Collins 1979b, Koster & Koet 1993, Prator 1968) should be addressed. This is also true of the inclusion of any suprasegmental errors – especially those of an intonational nature. More research needs to be done before it can be established

¹ This is a larger group than one may imagine, consisting not only of immigrants, exchange students, seasonal workers, or people whose partners, parents or relatives are speakers of these varieties, but also of those maintaining intensive professional contacts with these communities while resident in another country (including embassy staff and call centre workers).

² Teaching materials on Australian English pronunciation certainly exist, but Yates's (2001) survey of Australian teachers' attitudes to pronunciation teaching showed a need for more textbooks that deal specifically with this variety (Macdonald 2002).

that the incorrect use of weak and contracted forms, and certain intonational deviations, are as relatively insignificant as the analysis of the Native-speaker Experiment suggests. In this context, it would also be relevant to know to what extent the priority accorded to Dutch intonation by L1 speakers of English native speakers can be seen as more generally applicable to the intonation patterns of other groups of L2 speakers. An additional point for future investigation would be to find any confirmation for the suggestion that Dutch intonation patterns in American English may be perceived as “British”, and to discover whether this is perceived as a positive or a negative feature. A number of other features of the experiment (such as a three-syllable realisation of *secondary* or the spelling of *colour* with <ou>) were also described by some North American respondents as “British”, and consequently viewed as inappropriate. Research into this area could have significant consequences for English intonation teaching in the Netherlands.

Surprisingly, the results of the core experiment suggest that aspiration may not be an important acoustic cue for North Americans. This may be an interesting avenue of further enquiry. Similarly, it would be useful to gain more insight into Scottish and Northern Irish attitudes to non-native realisations of /ʊ, u:/, and to the conflation of these phonemes. This is especially important in view of the fact that these judges evaluated an example of this conflation more severely, in spite of the fact that they are unlikely to make such a contrast themselves (see 5.2.5). This was an interesting counter-example to the notion of “accent similarity” which deserves further investigation. Finally, given the fact that word stress was ranked so highly by all groups of respondents, it would seem expedient to investigate more closely the various phonetic and phonological factors which collectively determine the significance of word stress errors to native speakers of English.

6.2 Implications for teaching English in the Netherlands and elsewhere

6.2.1 Introduction

The results of the present study clearly indicate that learners of English should develop an awareness, both at a metalinguistic and a sociolinguistic level, of the different ways in which their English pronunciation is evaluated by various groups of native and non-native speakers (see 6.2.2 and 6.2.3 for practical suggestions on how this may be achieved). Not only will this provide them with a much more realistic appraisal of the effect of their L2 English accents on their interlocutors (both native and non-native), but it may also encourage them to monitor their own pronunciation and, if so desired, model it more closely on native-speaker English. In a world where, despite Jenkins’s asseverations, L1 speakers of English are still linguistically (if not numerically) dominant, this would empower L2 speakers of English to communicate more efficiently in

international contexts. A more native-like pronunciation will also enhance learners' ability to understand spoken English, since, as Trudgill (2005a) has argued, based on McAllister (1997), learners who cannot make certain phonemic distinctions will have greater difficulty perceiving these. In other words, L2 speakers of English should be *aware* of – rather than “beware” as Jenkins (2004) would have it – “the natives and their norms”. Needless to say, this can only be done by providing learners with pronunciation training at beginners', intermediate and advanced levels.

As the results of the Dutch Experiment have indicated, little or no attention is paid to pronunciation training for learners of English in the Netherlands until they reach higher education. On the basis of the findings from the core experiment, recommendations for such training will be made in 6.2.2 and 6.2.3 below. These have been divided into (1) basic skills to be taught at beginners' or intermediate level (e.g. in secondary schools), and (2) more advanced skills at higher levels (e.g. in universities and colleges). While this division reflects, to some extent, current pedagogical practices in the Netherlands, one could of course argue that detailed attention to pronunciation teaching should already be given in the initial stages of the curriculum, especially at an age when, in principle, learners are more receptive to this. However, this last is a moot point.

For instance, Scovel (1997: 119) has referred to the “simplistic and clearly mistaken notion that the earlier we introduce a foreign language to learners, the more fluent they will become”.³ Furthermore, Bongaerts (1999b) provides examples of early learners who still developed a foreign accent, and late learners who became indistinguishable from native speakers.⁴ Similarly, research by Guion *et al.* (2004: 38) into the acquisition of English word stress by early and late Spanish bilinguals rejects the notion that “learners older than 12 would not be able to acquire phonological knowledge in a second language and learners younger than seven would be native-like in this knowledge”.⁵ Very recent findings also suggest that adults can be trained more successfully to make phoneme distinctions in L2 speech than was previously assumed (Brew 2005). In view of these conflicting considerations, the possibly contentious division

³ It should be noted that in an earlier study, Scovel (1988: 122) specifically postulated a “critical period” for speech only.

⁴ For instance, Bongaerts (1999b: 136) refers to research by Flege *et al.* (1995) which showed that no fewer than 22% of English-speaking Italian respondents living in Canada who had begun to learn English at a young age failed to develop a native accent in that language. In the same study, Bongaerts (1999b: 154) also describes three experiments carried out by himself and his co-workers which revealed that a number of very advanced L1 Dutch learners of English and French did in fact achieve a “native-like” or “authentic” pronunciation of those languages as adults.

⁵ This is based on two experiments conducted by Guion *et al.* (2004), in which it was tested whether early and late L1 Spanish learners of US English can acquire a native-like ability to produce and predict English stress patterns accurately in a number of non-words.

into two levels has been retained. In fact, if no changes are made to the present situation, whereby little or no pronunciation training is provided in secondary schools, most or all aspects of the curricula suggested in 6.2.2 and 6.2.3 will have to be covered in higher education. It is already the case that universities and colleges in the Netherlands, in addition to raising students' awareness of pronunciation issues, find themselves having to provide very basic training in this subject.

Administrators and pedagogues responsible for developing and implementing English curricula in higher education should note the low priority given to pronunciation training in secondary schools. If they are convinced that non-distracting accents will empower university and college graduates to communicate efficiently with both native and non-native speakers of English, they should ensure that pronunciation training continues to be securely embedded in their programmes. In particular, teacher training colleges would appear to have a responsibility to make sure that prospective teachers have reasonably convincing accents in English. As is stated in Gimson & Cruttenden (1994: 273), the "foreign teacher of English ... has the obligation to present his students with as faithful a model of English pronunciation as is possible", if only – as they point out – because his or her accent may be imitated by their pupils. Those engaged in teaching or administration in higher education have a duty to establish co-operation between schools and universities in order to encourage pronunciation training at the level of secondary education. Without this, there will be a constant need to help undergraduate students of English unlearn Dutch-influenced pronunciation habits unconsciously acquired in prior educational settings.

6.2.2 Beginners' or intermediate level

Apart from the finding that secondary school teachers of English appear to pay little or no attention to pronunciation training, the present study has also revealed that a small minority do not employ the medium of English at all. If it is an important objective of English language teaching in the Netherlands to empower learners to communicate effectively in exchanges with other non-Dutch speakers of English, this clearly implies that both English-medium instruction and English pronunciation training should be part of the standard curriculum.

An additional reason to give more emphasis to pronunciation training is that pupils' accents may be subject to evaluation in all years. Not only is this clear from the results of the Dutch Experiment, but it is also evident from the English syllabus for pre-university education as approved by the Dutch Department of Education. According to this syllabus, pupils' fluency should be evaluated on the basis of a number of criteria which include pronunciation (Examenblad 2005: Domein C). It is even in accordance with the less ambitious objectives formulated for foreign-language teaching in the lower streams of secondary education, which require pupils to have "a certain degree of correctness" in their pronunciation (Ministerie van OCW 2003: 20, present author's

translation). It would be unreasonable to evaluate learners' accents without providing them with any prior training in this.

If pronunciation training is to be more integrated into the curriculum, it would be useful to investigate first which factors have led to this subject being neglected. As participants' comments in the Dutch survey indicate, such factors may include overburdened programmes, unrealistic class sizes, together, possibly, with a desire not to tax pupils unnecessarily. Avoidance of pronunciation teaching may also be linked to teachers' lack of training in this area (see Morley 1996) or to negative attitudes to the subject. As a result of such "deficit model" approaches to pronunciation training, the subject is not taught, because it is perceived as being demoralising to learners (see Yates 2001).

For pronunciation training to be effective, it is important that learners be exposed to a native-speaker model which they feel motivated to imitate. This is a further reason as to why it is desirable for English school teachers in the Netherlands to have a reasonably convincing command of a native accent of English themselves. The results of the Dutch Experiment suggest that this is not always the case. Teachers should make more explicit reference to different pronunciation models such as RP and GA (without of course prejudicing learners against any particular variety), and should discuss a number of salient differences between the sound systems of Dutch, RP, GA or other relevant models. This is particularly relevant given the fact that the media currently provide Dutch learners with an unprecedented level of exposure to different varieties of English; arguably, learners would be more encouraged to relate this to their English classes if they were presented with a structural metalinguistic framework to deal with such linguistic variation. Learners might also find it easier to evaluate and possibly improve their own pronunciation if they were aware of the target models against which their progress could be measured.

The results of the Native-speaker Experiment suggest that judges considered a number of Dutch pronunciation problems to be of significance. If one considers the errors presented in Table 5.1 (see 5.2.2), the most serious are likely to be those with severity estimates over 2.2 Likert scale points. These have been included in a list of urgent pronunciation problems to be addressed at more basic level (see Table 6.1). Apart from word stress and stress-related errors, these comprise not only phonemic errors but also "sub-phonemic" realisations that are likely to confuse or irritate native speakers. It should be noted that some of these, such as uvular-*r*, schwa-insertion and inappropriate *r*-deletion, are in fact more likely to be found in speakers of varieties of Dutch other than ABN (see Collins & Mees 2003: 179, 198, 201). As a result, there will be little reason to practise these errors with learners unless they are a feature of their accents. In addition, Table 6.1 does not include any pronunciation problems that are more typical of advanced learners, such as the glottalisation of lenis stops (Collins & Mees 2003: 153). Furthermore, it should be noted that this list is based on native-speaker judgements rather than on any other pedagogical considerations. Even if inappropriate use of post-vocalic *r* was considered less significant by participants in the RP version than was *r*-deletion by their GA

counterparts, the effect of **r**-deletion on the realisation of RP vowels such /ɑ: ɜ: ə ɔ:/ and a number of diphthongs may still warrant the inclusion of this notorious problem area (see Collins & Mees 2003: 180–181). It is also helpful for pupils to be aware of phenomena such as **r**-deletion and **r**-insertion, as these are salient indicators of differences between major accents.

Table 6.1. Overview of pronunciation problems to be urgently addressed at a more basic level, arranged by target accent (RP or GA).

Received Pronunciation	General American
<p>Word stress and related errors</p> <p>Phonemic errors:</p> <ul style="list-style-type: none"> • Fortis/lenis neutralisation • Conflation of /æ ~ e, ʌ ~ ɒ, ʊ ~ u:/ and /v ~ w/ • Some substitutions of /θ, ð/ by /t, d/ <p>Realisations leading to phoneme conflation:</p> <ul style="list-style-type: none"> • Lack of aspiration in initial stops <p>Realisations leading to irritation:</p> <ul style="list-style-type: none"> • Use of uvular-r • Schwa-insertion in /lm/ coda clusters 	<p>Word stress and related errors</p> <p>Phonemic errors:</p> <ul style="list-style-type: none"> • Fortis/lenis neutralisation • Conflation of /æ ~ e/ and /v ~ w/ • Most substitutions of /θ, ð/ by /t, d/ <p>Realisations leading to irritation:</p> <ul style="list-style-type: none"> • Use of uvular-r • Schwa-insertion in /lm/ coda clusters • Inappropriate r-deletion

Finally, it should be pointed out that the very high severity estimates given to word stress errors suggest that this should be a major area of concern in pronunciation teaching (as is also proposed by Gussenhoven & Broeders 1997: 16). Many teachers of English in the Netherlands will be familiar with the widespread tendency to stress incorrectly certain high-frequency items such as **Arabic*, **Catholicism*, **development*, **orchestra*, **politics* or **var[ai]able*, to cite a few of many examples, even in otherwise advanced learners. It is a well-known fact that the Dutch and English word stress “systems” have a high degree of similarity (Trommelen & Zonneveld 1999), but native-speaker judges clearly also attach a great deal of importance to the correct pronunciation of those words which deviate from the basic pattern (whether idiosyncratically so or because of the influence of morphological structure, as in these examples). If correct word stress is viewed as such a strong indicator of the L2 English speaker’s successful mastery of the language, this argues in favour of a central position for word stress in pronunciation training. Moreover, it should be noted that stress will typically have significant effects on the realisation of vowels, resulting in vowel

gradation. As a consequence, it can be said that stress errors almost always imply phonemic errors. (For different accounts of the role of vowel gradation in recognising stressed and unstressed syllables, see Fear *et al.* 1995 and Mattys 2000.)

The results of the present study imply that teachers should only evaluate learners' pronunciation by the standards of the target accent in question. For RP, this means that teaching learners to avoid /ʌ ~ ʊ/ confusion is more important than warning them off certain "Americanisms" such as yod-deletion in *new*. For GA, this implies, for instance, that substituting /θ, ð/ by /t, d/ is almost always considerably more serious than /ʌ ~ ʊ/ confusion. It may help to point out to learners that TH-stopping is subject to stigmatisation, especially in American English. This is a useful corrective to the notion anecdotally attested for some Dutch learners of English that their strong L2 accents are somehow or other more "American". This evasive "manoeuvre" on the part of advanced pronunciation learners has frequently been observed by the present author within the context of pronunciation training given to first-year and second-year students of English at the Universities of Utrecht and Leiden as part of the proficiency curriculum. In fact, as the findings of the present study suggest, considerable caution is required in encouraging learners to retain Dutch English pronunciation features merely because these are similar to what is heard in some native-speaker varieties.

6.2.3 Advanced level

Apart from the more basic pronunciation problems discussed in 6.2.2, advanced learners, such as those at universities and other kinds of higher education, should make themselves aware of the less serious errors listed as in Table 5.1 (see 5.2.2). This implies that in addition to word stress and certain phonemic and stigmatised errors, they should also pay considerable attention to phonetic detail, especially, but not exclusively, in those cases where particular phonetic realisations may lead to phoneme conflation. At this level, it is notably the use of glottal stops for /d/ that should be given high priority, particularly since this error is typical of more advanced Dutch learners. Another pronunciation problem to be prioritised more clearly and effectively is the English constraint on degemination. Learners may also benefit from an awareness of the effects of phonological context, and of word frequency, on error severity (as a result of which, for instance, substituting /ð/ by /d/ is more serious in *together* than in *that*). Moreover, learners should also practise suprasegmental features such as vowel gradation and appropriate intonation. This is a well-established practice, considered essential in many textbooks (see Gimson & Cruttenden 1994: 276, 281, 287), which should not simply be discontinued because of the inconclusive results yielded in this respect by the present experiment.

The hierarchies of error for Dutch learners of RP and GA as represented in Table 5.1 largely correspond to those proposed in Collins & Mees (2003: 290–291, 1993: 124–130). This implies that these hierarchies can continue to be used in pronunciation training at Dutch universities and colleges. However,

the results of the Native-speaker Experiment show that some errors were less salient, or were evaluated less seriously, than has previously been assumed. Especially for learners of RP or similar British accent models, these include the use of overdark pharyngealised [ɮ] and, possibly, overlong /aɪ/. For learners of North American accents, and GA in particular, aspiration of initial stops may be less of a priority, as is true of the conflation of /ʌ ~ ɒ, ʊ ~ u:/. The fact that certain pronunciations may be stigmatised in one variety, but less so or not at all in another, demonstrates that pronunciation trainers and learners need to distinguish clearly between these different accent models, as Preisler (1999: 264) has also advocated for Danish learners of English. This reinforces the well-known injunction to advanced learners of English to avoid “dialect mixing” or “Mid-Atlantic English” and to adhere to one particular model as consistently as possible.⁶ It also argues against the adoption of a “Lingua Franca Core” (Jenkins 2000) in pronunciation teaching in the Netherlands.

It is important to stress that advanced learners will also benefit from a metalinguistic awareness of accent variation in English, and sociolinguistic sensitivity to this subject. Such awareness could be fostered as part of learners’ pronunciation training. For instance, it is to be strongly recommended that learners be made aware of the stigma attached to certain pronunciation features in different groups of native speakers. This accounts for the particularly severe evaluations in North America of errors such as fortis/lenis neutralisation, TH-stopping and deletion or weakening of **r** and **l**. Such stigmatisation was also attested in speakers of regional accents, even with regard to features directly associated with these accents. Awareness of such attitudes will be useful in helping students avoid strongly marked realisations (cf. Morley 1996: 149). However, training of this nature will also require a certain level of sensitivity and sociolinguistic insight on the part of the instructors and their students. As Swacker (1976: 16) has pointed out, “[t]he task of the linguistically sophisticated foreign language instructor is bifurcated in the most awkward way in that we must simultaneously strive to develop good attitudes of dialect tolerance in our students while guiding them into dialectal patterns that will best facilitate their widest acceptance into a community of target language speakers”.

⁶ For instance, Gimson & Cruttenden (1994: 276) warn against the use of American speech forms if the learner’s model is RP. Consistent adherence to one particular model is also one of the criteria used in evaluating students’ pronunciation in English degree courses co-taught by the present author at the Universities of Utrecht and Leiden for over a decade. However, Van der Haagen (1998: 105) has suggested that secondary school teachers should allow their pupils to speak a type of English that “sometimes follows the rules for RP and sometimes those for GA”, even though she does not extend the same latitude to teachers, since “parents and pupils do not expect teachers to have a Mid-Atlantic accent”. Nevertheless, Modiano (1996) has propagated a “Mid-Atlantic” model of English for L2 learners, as a compromise between British and American English, where “marked” forms which are not “internationally” intelligible are eschewed. It turns out, however, that the pronunciation features described as “marked” are without exception associated with British English.

Such sociolinguistic sensitivity and competence may well be an essential aspect of the language acquisition process for very advanced learners (see Bayley & Regan 2004).

As Trudgill (2005b: 84) has argued, “the most linguistically offended against and negatively evaluated English speakers of all are undoubtedly the vast majorities of the populations of Britain, North America and Australasia who speak native but non-standard varieties of the language”. While it is in the interest of what one may term “accent learners” to be warned against such negative evaluations, this does not of course mean that they should be encouraged to internalise any of these prejudices against non-standard accents, or their speakers. Nor does it mean that learners should insist that speakers of regional accents “modify” their accents “in an international setting”, as Jenkins (2000: 228) actually suggests – quoting a request from a Japanese learner to native speakers to “drop the dialects”. Clearly, it would be socially unacceptable for a non-native speaker of English to demand that a speaker of Irish English, or of AAVE, modify their accent in the direction of the standard favoured by the learner.⁷ What this does imply, however, is that advanced learners should only use attitudinally marked forms if they have a strongly integrative motivation to learn the variety with which these realisations are associated. Even then, they should be aware, as Swacker (1976: 16) has put it, that “the native speaker is quite ready to reject from the foreign speaker exactly those regional markers he personally identifies with in his own speech”.

⁷ One wonders how different this is from a learner of Dutch requiring a speaker of a broad Flemish accent to modify this in the direction of Standard Dutch. In a slightly different context, it may even be compared to a learner of Spanish asking a Catalan speaker to switch to Castilian.

REFERENCES

- Abbott, G. (1986). A new look at phonological 'redundancy'. *ELT Journal* **40**. 299–305. Reprinted in Brown (1991). 225–234.
- Abercrombie, D. (1956). Teaching pronunciation. In D. Abercrombie *Problems and principles: studies in the teaching of English as a second language*. London: Longmans. 28–40. Reprinted in Brown (1991). 89–95.
- Aceto, M. (2004). Eastern Caribbean English-derived language varieties: phonology. In Schneider *et al.* (2004). 481–500.
- Aitken, A. J. (1984). Scottish accents and dialects. In Trudgill (1984b). 94–118.
- Albrechtsen, D., B. Henriksen & C. Færch (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning* **30**. 365–396.
- Algeo, J. (ed.) (2001). *The Cambridge history of the English language*. Volume 6: *English in North America*. Cambridge: Cambridge University Press.
- Altendorf, U. & D. Watt (2004). The dialects in the south of England: phonology. In Schneider *et al.* (2004). 178–203.
- Anderson, V. (2001). Devoiced obstruents in Pennsylvania Dutchified English: an OT analysis. Paper presented at the 7th meeting of the Mid-Continental Workshop on Phonology, University of Iowa.
- Anderson, V. (2002). Below the threshold of perception: change in devoiced obstruents in Pennsylvania Dutchified English. Paper presented at the 31st Conference on New Ways of Analyzing Variation, Stanford University.
- Anderson-Hsieh, J., R. Johnson & K. Koehler (1992). The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* **42**. 529–555.
- Anderson-Hsieh, J. & K. Koehler (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning* **38**. 561–613.
- Ash, S. (1982). The vocalization of intervocalic /l/ in Philadelphia. *SECOL Review* **6**. 162–175.
- Bansal, R. K. (1965/66). *The intelligibility of Indian English: measurements of the intelligibility of connected speech and sentence and word material, presented to listeners of different nationalities*. PhD dissertation, University College, London.
- Bansal, R. K. (1969). *The intelligibility of Indian English*. Hyderabad: Central Institute of English.
- Baskaran, L. (2004). Malaysian English: phonology. In Schneider *et al.* (2004). 1034–1046.
- Bauer, L. (1994). English in New Zealand. In Burchfield (1994a). 382–429.
- Bauer, L. (2002). *An introduction to international varieties of English*. Edinburgh: Edinburgh University Press.

- Bauer, L. & P. Warren (2004). New Zealand English: phonology. In Schneider *et al.* (2004). 580–602.
- Bayley, R. (2000). In addition to English: second-language acquisition and variationist linguistics. *American Speech* **75**. 288–290.
- Bayley, R. & V. Regan (2004). Introduction: the acquisition of sociolinguistic competence. *Journal of Sociolinguistics* **8**. 323–338.
- Beal, J. (2004). English dialects in the north of England: phonology. In Schneider *et al.* (2004). 113–133.
- Beckman, J. N. (1999). *Positional faithfulness: an optimality theoretic treatment of phonological asymmetries*. Garland: New York.
- Bernard, M., L. Bonnie, S. Riley, T. Hackler & K. Hanzen (2002). A comparison of popular online fonts: which size and type is best? *Usability News* **4:1**. psychology. wichita.edu/surl/usabilitynews/41/onlinetext.htm. Accessed 16 May 2006.
- Bex, T. & J. Watts (eds.) (1999). *Standard English: the widening debate*. London: Routledge.
- Boberg, C. (2004). English in Canada: phonology. In Schneider *et al.* (2004). 351–365.
- Boersma, P. & D. Weenink (2002). Praat: doing phonetics by computer (Version 4.0.9). www.praat.org. Accessed 1 March 2002.
- Bongaerts, T. (1999a). De keuze van beoordelaars in onderzoek naar uitspraakvaardigheid in een vreemde taal. In M. Gerritsen & D. Springorum (eds.) *Bedrijfscommunicatie: een bundel voor Ger Peerbooms bij gelegenheid van zijn 65^e verjaardag*. Nijmegen: Nijmegen University Press. 1–10.
- Bongaerts, T. (1999b). Ultimate attainment in L2 pronunciation: the case of very advanced late L2 learners. In D. Birdsong (ed.) *Second language acquisition and the critical period hypothesis*. Mahwah: Erlbaum. 133–159.
- Bongaerts, T., S. Mennen & F. van der Slik (2000). Authenticity of pronunciation in naturalistic second language acquisition: the case of very advanced late learners of Dutch as a second language. *Studia Linguistica* **54**. 298–308.
- Booij, G. E. (1977). *Dutch morphology: a study of word formation in generative grammar*. Dordrecht: Foris.
- Booij, G. E. (1995). *The phonology of Dutch*. Oxford: Clarendon Press.
- Bot, C. de (1982). *Visuele feedback van intonatie*. PhD dissertation, Nijmegen University. Published, Enschede: Sneldruk Boulevard.
- Bowerman, S. (2004). White South African English: phonology. In Schneider *et al.* (2004). 931–942.
- Bradley, N. (1999). Sampling for Internet surveys: an examination of respondent selection for Internet research. *Journal of the Market Research Society* **41**. 387–395.
- Branford, W. (1994). English in South Africa. In Burchfield (1994a). 430–496.
- Bremmer, R. H. & C. Gussenhoven (1983). Voiced fricatives in Dutch: sources and present-day usage. *North-western European Language Evolution* **2**. 55–71.

- Bresnahan, M. J., R. Ohashi, R. Nebashi, W. Y. Liu & S. M. Shearman (2002). Attitudinal and affective response toward accented English. *Language and Communication* **22**. 171–185.
- Brew, A. (2005). Adults can be retrained to learn second languages more easily, says UCL scientist. *Eurekalert!* 14 June 2005. www.eurekalert.org/pub_releases/2005-06/potn-acb061405.php. Accessed 16 May 2006.
- Brinton, L. J. & M. Fee (2001). Canadian English. In Algeo (2001). 422–440.
- Broatch, M. (2002). E-tales: not a Whittaker's bar in sight. *Computerworld* 15 July 2002. computerworld.co.nz/cw.nsf/unid/cc256d400014e76ccc256bf20074618?opendocument&highlight=2,doel. Accessed 16 May 2006.
- Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly* **22**. 593–606. Reprinted in Brown (1991). 211–224.
- Brown, A. (ed.) (1991). *Teaching English pronunciation: a book of readings*. London: Routledge.
- Bruthiaux, P. (2003). Squaring the circles: issues in modeling English worldwide. *International Journal of Applied Linguistics* **13**. 159–178.
- Burchfield, R. (ed.) (1994a). *The Cambridge history of the English language*. Volume 5: *English in Britain and overseas: origins and development*. Cambridge: Cambridge University Press.
- Burchfield, R. (1994b). Introduction. In Burchfield (1994a). 1–19.
- Cassidy, F. C. (ed.) (1985a). *Dictionary of American regional English*. Volume 1: *Introduction and A–C*. Cambridge, Mass.: Belknap.
- Cassidy, F. C. (1985b). Language changes especially common in American folk speech. In Cassidy (1985a). xxxvi–xl.
- Cassidy, F. C. (ed.) (1991). *Dictionary of American regional English*. Volume 2: *D–H*. Cambridge, Mass.: Belknap.
- Cassidy, F. C. & J. Houston Hall (eds.) (1996). *Dictionary of American regional English*. Volume 3: *I–O*. Cambridge, Mass.: Belknap.
- Chastain, K. (1980). Native speaker reaction to instructor-identified student second-language errors. *Modern Language Journal* **64**. 210–215.
- Christophersen, P. (1973). *Second language learning: myth and reality*. Penguin: Harmondsworth.
- Clark, U. (2004). The English West Midlands: phonology. In Schneider *et al.* (2004). 134–162.
- Clarke, S. (1993). The Americanization of Canadian pronunciation: a survey of palatal glide usage. In S. Clarke (ed.) *Focus on Canada*. Amsterdam & Philadelphia: John Benjamins. 85–108.
- Clarke, S. (2004). Newfoundland English: phonology. In Schneider *et al.* (2004). 366–382.
- Clayton, J. F. (2004). Using the Internet to collect quantitative data. www.wintec.ac.nz/files/about%20us/services/clt/withit/volume3/itpnz_clayton.doc. Accessed 16 May 2006.
- Clyne, M. G. (1992). *Pluricentric languages: differing norms in different nations*. Berlin: Mouton de Gruyter.

- Collins, B. (1979a). Hierarchy of error of Dutch learners. Paper presented at the 2nd International Conference on the Teaching of Spoken English, University of Leeds.
- Collins, B. (1979b). The significance of pronunciation errors. *IATEFL Newsletter* **56**. 26–29.
- Collins, B., S. P. den Hollander, I. M. Mees & J. Rodd (2001). *Sounding better: a practical guide to English pronunciation for speakers of Dutch*. Holten: Walvaboek.
- Collins, B., S. P. den Hollander & J. Rodd (1987). *Accepted English pronunciation*. Apeldoorn: Van Walraven.
- Collins, B. & I. M. Mees (1981). *The sounds of English and Dutch*. The Hague: Leiden University Press.
- Collins, B. & I. M. Mees (1993). *Accepted American pronunciation*. Apeldoorn: Van Walraven.
- Collins, B. & I. M. Mees (2003a). *Practical phonetics and phonology: a resource book for students*. London & New York: Routledge.
- Collins, B. & I. M. Mees (2003b). *The phonetics of English and Dutch*. 5th revised edn. Leiden: Brill.
- Corder, S. P. (1973). *Introducing applied linguistics*. Harmondsworth: Penguin.
- Cote, P. & R. Clement (1994). Language attitudes: an interactive situated approach. *Language and Communication* **14**. 237–252.
- Couper, M. P. (2000). Web surveys: a review of issues and approaches. *Public Opinion Quarterly* **64**. 464–481.
- Cruttenden, A. (1997). *Intonation*. Cambridge: Cambridge University Press.
- Crystal, D. (2001). The future of Englishes. In A. Burns & C. Coffin (eds.) *Analysing English in a global context*. London: Routledge. 53–64.
- Cunningham-Andersson, U. (1997). Native speaker reactions to non-native speech. In Leather & James (1997). 133–144.
- Cutler, A., D. Dahan & W. van Donselaar (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech* **40**. 141–201.
- Cutler, A., A. Weber, R. Smits & N. Cooper (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* **116**. 3668–3678.
- Dalton, C. & B. Seidlhofer (1994). *Pronunciation*. Oxford: Oxford University Press.
- Daniels, H. (1995). Psycholinguistic, psycho-affective and procedural factors in the acquisition of authentic L2 phonology. In D. Hill (ed.) *Bologna '94 English language teaching*. Milan: British Council. 77–82. Reprinted in M. Vaughan Rees (ed.) (1995). *Speak Out!* **15**. 3–10 and in A. McLean (ed.) (1997). *SIG selections 1997: Special interests in ELT*. Whitstable: IATEFL. 80–85.
- Davenport, M. & S. J. Hannahs (1998). *Introducing phonetics and phonology*. London & New York: Arnold & Oxford University Press.
- Davies, E. (1983). Error evaluation: the importance of viewpoint. *ELT Journal* **37**. 304–311.

- Delamare, T. (1996). The importance of interlanguage errors with respect to stereotyping by native speakers in their judgements of second language learners' performance. *System* **24**. 279–297.
- Devonish, H. & O. G. Harry (2004). Jamaican Creole and Jamaican English: phonology. In Schneider *et al.* (2004). 450–480.
- Dillman, D., R. Tortora & D. Bowker (1998). Principles for constructing web surveys. *SESRC Technical Report* **98**. survey.sesrc.wsu.edu/dillman/papers/websurveyppr.pdf. Accessed 16 May 2006.
- Dretzke, B. (1985). *Fehlerbewertung im Aussprachebereich: objective Fehlerbeurteilung versus subjective Fehlerbewertung: eine Untersuchung von Aussprachefehlern deutscher Anglistikstudenten in der Zielsprache English*. Hamburg: Buske.
- Dubois, S. & B. M. Horvath (2004). Cajun Vernacular English: phonology. In Schneider *et al.* (2004). 407–416.
- Dulay, H., M. Burt & S. Krashen (1982). *Language two*. Oxford: Oxford University Press.
- Dziubalska-Kołodziejczyk, K. & J. Przedlacka (eds.) (2005). *English pronunciation models: a changing scene*. Berne: Peter Lang.
- Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning* **27**. 315–330. Reprinted in G. Ioup & S. Weinberger (eds.) (1987). *Interlanguage phonology*. Cambridge, Mass.: Newbury House. 55–69.
- Edwards, W. F. (2004). African American Vernacular English: phonology. In Schneider *et al.* (2004). 383–392.
- Eisenstein, M. (1983). Native reactions to non-native speech: a review of empirical research. *Studies in Second Language Acquisition* **5**. 160–176.
- Ellis, A. (1889). *On early English pronunciation*. Part 5: *Existing dialectal as compared with West Saxon pronunciation*. London: Early English Text Society.
- Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.
- European Commission (2005). Europeans and languages. *Special Eurobarometer* **237**. www.europa.eu.int/comm/public_opinion/archives/ebs/ebs_237.en.pdf. Accessed 16 May 2006.
- Examenblad (2005). Bijlage 3: examenprogramma Engelse taal en letterkunde v.w.o. www.examenblad.nl/9336000/1/j9vvgodkvkzp4d4/vg41h1jsjgxx/f=/bestand.doc. Accessed 16 May 2006.
- Fayer, J. M. & E. Krasinski (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning* **37**. 313–325.
- Fear, B. D., A. Cutler & S. Butterfield (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America* **97**. 1893–1904.
- Flege, J., M. Munro & R. MacKay (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America* **97**. 3125–3134.

- Fougeron, C. & P. A. Keating (1997). Articulatory strengthening at the edges of prosodic domains. *Journal of the Acoustical Society of America* **101**. 3728–3740.
- Foulkes, P. & G. Docherty (eds.) (1999). *Urban voices: accent studies in the British Isles*. London: Arnold.
- Frary, R. B. (1996). Hints for designing effective questionnaires. *Practical Assessment, Research & Evaluation* **5**. www.cmu.edu/teaching/assessment/resources/SurveyGuidelines.pdf. Accessed 16 May 2006.
- Fraser, B. (2002). Re: English with an accent. 18 July 2002. groups.google.com/group/soc.culture.scottish/browse_thread/thread/70af95690f15fdf0/77e1be2a5c4f9abf?q=doel+accent&rnum=1#77e1be2a5c4f9abf. Accessed 16 May 2006.
- Gaddis, S. E. (1998). How to design online surveys. *Training and Development* **52**. 67–71.
- Galloway, V. B. (1980). Perceptions of the communicative effects of errors in Spanish. *Modern Language Journal* **64**. 428–453.
- Gass, S. & E. M. Varonis (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* **34**. 65–89.
- Gibbon, D. (2005). Afterword. In Dziubalska-Kołaczyk & Przedlacka (2005). 439–465.
- Gick, B. (2002). The American intrusive L. *American Speech* **77**. 167–183.
- Giles, H. (1970). Evaluative reactions to accents. *Educational Review* **22**. 211–227.
- Giles, H. (1971). Patterns of evaluation in reaction to RP, South Welsh and Somerset accented speech. *British Journal of Social and Clinical Psychology* **10**. 280–281.
- Giles, H. & P. F. Powesland (1975). *Speech style and social evaluation*. London: Academic Press.
- Gimson, A. C. (1978). Towards an international pronunciation of English. In P. Stevens (ed.) *In honour of A. S. Hornby*. Oxford: Oxford University Press. 45–53.
- Gimson, A. C. & A. Cruttenden (1994). *Gimson's pronunciation of English*. 5th edn. London: Arnold.
- Glenallan (2002). Re: English with an accent. 19 July 2002. groups.google.com/group/soc.culture.scottish/browse_thread/thread/70af95690f15fdf0/77e1be2a5c4f9abf?q=doel+accent&rnum=1#77e1be2a5c4f9abf. Accessed 16 May 2006.
- GoldWave Digital Audio Editor* (2002). Version 4.26. www.goldwave.com. Accessed 1 April 2002.
- Gordon, E. & M. Maclagan (2004). Regional and social differences in New Zealand: phonology. In Schneider *et al.* (2004). 603–613.
- Gordon, M. J. (2004a). New York, Philadelphia, and other northern cities: phonology. In Schneider *et al.* (2004). 282–299.
- Gordon, M. J. (2004b). The West and Midwest: phonology. In Schneider *et al.* (2004). 338–350.

- Guion, S. G., T. Harada & J. J. Clark (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. Pre-publication version available at uoregon.edu/~guion/Guion%20et%20al.%20for%20BLC.pdf. Accessed 16 May 2006. Published in *Bilingualism: Language and Cognition* 7. 207–226.
- Gunn, H. (2002). Web-based surveys: changing the survey process. *First Monday* 7. www.firstmonday.org/issues/issue7_12/gunn/. Accessed 16 May 2006.
- Guntermann, G. (1978). A study of the frequency and communicative effects of errors in Spanish. *Modern Language Journal* 62. 249–253.
- Gussenhoven, C. & A. Broeders (1997). *English pronunciation for student teachers*. 2nd edn. Groningen: Wolters Noordhoff.
- Gussenhoven, C., T. Rietveld, J. Kerkhoff & J. M. B. Terken (2003). ToDI: Transcription of Dutch Intonation. 2nd edn. todi.let.kun.nl/todi/home.htm. Accessed 16 May 2006.
- Haagen, M. van der (1998). *Caught between norms: the English pronunciation of Dutch learners*. PhD dissertation, Nijmegen University. Published, The Hague: Holland Academic Graphics.
- Harris, J. (1994). *English sound structure*. Oxford: Blackwell.
- Hartman, J. W. (1985). Guide to pronunciation. In Cassidy (1985a). xlii–lxi.
- Hay, J. & A. Sudbury (2005). How rhoticity became /r/-sandhi. *Language* 81. 799–823.
- Hickey, R. (1999). Dublin English: current changes and their motivation. In Foulkes & Docherty (1999). 265–281.
- Hickey, R. (2004a). *A sound atlas of Irish English*. Berlin: Mouton de Gruyter. [Includes DVD].
- Hickey, R. (2004b). Irish English: phonology. In Schneider *et al.* (2004). 68–97.
- Higgs, J. (1980). The American /r/ is advanced velar not post-alveolar! *Work in Progress* 13. Edinburgh: Department of Linguistics, Edinburgh University.
- Holliday, A. (2005). *The struggle to teach English as an international language*. Oxford: Oxford University Press.
- Horvath, B. M. (2004). Australian English: phonology. In Schneider *et al.* (2004). 625–644.
- Houston Hall, J. (ed.) (2002). *Dictionary of American regional English*. Volume 4: P–Sk. Cambridge, Mass.: Belknap.
- Hughes, A. & C. Lascaratou (1982). Competing criteria for error gravity. *ELT Journal* 36. 175–182.
- Hughes, A. & P. Trudgill (1987). *English accents and dialects: an introduction to social and regional varieties of British English*. 2nd edn. London: Arnold.
- Ihalainen, O. (1994). The dialects of English since 1776. In Burchfield (1994a). 1–19.
- Information Society Promotion Office of the European Commission (1999). Alternative Networks: Netherlands. *ESIS Knowledge Base*. www.eu-esis.org/Alternative/NLaltQ8.htm. Accessed 16 May 2006.

- Jenkins, J. (2000). *The phonology of English as an international language: new models, new norms, new goals*. Oxford: Oxford University Press.
- Jenkins, J. (2004). Beware the natives and their norms. *Guardian Weekly* 22 January 2004. education.guardian.co.uk/tefl/story/0,,1128998,00.html. Accessed 16 May 2006.
- Jenner, B. (1989). Teaching pronunciation: the common core. *Speak Out!* 4. 2–4.
- Johansson, S. (1973). The identification and evaluation of errors in foreign languages: a functional approach. In J. Svartvik (ed.) *Errata: papers in error analysis*. Lund: Gleerup. 102–114.
- Johansson, S. (1975). *Papers in contrastive linguistics and language testing*. Lund: Gleerup.
- Johansson, S. (1978). *Studies in error gravity: native reactions to errors produced by Swedish learners of English*. Gothenburg: Acta Universitatis Gothoburgensis.
- Johnson, K. (2005). Speaker normalization in speech perception. In Pisoni & Remez (2005). 363–389.
- Jones, D. (2003). *English pronouncing dictionary*. 16th edn, edited by P. Roach, J. Hartman & J. Setter. Cambridge: Cambridge University Press.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk & H. G. Widdowson (eds.) *English in the world: teaching and learning the language and literatures*. Cambridge: Cambridge University Press. 11–30.
- Kenyon, J. S. & T. A. Knott (1953). *A pronouncing dictionary of American English*. Springfield, Mass.: Merriam-Webster.
- Kingston, J. (1985). *The phonetics and phonology of the timing of oral and glottal events*. PhD dissertation, University of California, Berkeley.
- Kingston, J. (1990). Articulatory binding. In J. Kingston & M. Beckman (eds.) *Papers in laboratory phonology*. Volume 1: *Between the grammar and physics of speech*. Cambridge: Cambridge University Press. 406–434.
- Kirwin, W. J. (2001). Newfoundland English. In Algeo (2001). 253–290.
- Klaassen, R. (2002). The international university curriculum: challenges in English-medium engineering education. *Landelijk Overleg Studievaardigheden Contact* 22. 24–26. www-dsz.service.rug.nl/los/LOSCON/Nr22/AF2/iuc.htm. Accessed 16 May 2006.
- Knowles, G. (1992). Stress. In McArthur (1992a). 988–989.
- Koster, C. J. & T. Koet (1993). The evaluation of accent in the English of Dutchmen. *Language Learning* 43. 69–92.
- Kreft, I. & J. de Leeuw (1998). *Introducing multi-level modeling*. London: Sage Publications.
- Labov, W. (1966). *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
- Labov, W. (1991). The three dialects of English. In P. Eckert (ed.) *New ways of analyzing sound change*. San Diego: Academic Press. 1–44.

- Labov, W. (2001). *Principles of linguistic change*. Volume 2: *Social factors*. Oxford: Blackwell.
- Labov, W., S. Ash & C. Boberg (2005a). The atlas of North American English: Map 6. www.ling.upenn.edu/phono_atlas/maps/Map6.html. Accessed 8 July 2005. To appear in Labov *et al.* (in press).
- Labov, W., S. Ash & C. Boberg (2005b). The atlas of North American English: TELSUR map IN-3. www.ling.upenn.edu/phono_atlas/maps/MapsIN/TelsurIN_uh.html. Accessed 8 July 2005. To appear in Labov *et al.* (in press).
- Labov, W., S. Ash & C. Boberg (in press). *The atlas of North American English: phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.
- Ladegaard, H. J. (1998). National stereotypes and language attitudes: the perception of British, American and Australian language and culture in Denmark. *Language and Communication* **18**, 251–274.
- Lambert, W. E. (1967). A social psychology of bilingualism. *Journal of Social Issues* **23**, 91–108. Reprinted in J. B. Pride & J. Holmes (1972). *Sociolinguistics: selected readings*. Penguin: Harmondsworth. 336–349.
- Lance, D. M. (1999). Regional variation in subjective dialect divisions in the United States. In Preston (1999a). 283–314.
- Lass, R. (1990). A ‘standard’ South African vowel system. In S. Ramsaran (ed.) *Studies in the pronunciation of English: a commemorative volume in honour of A. C. Gimson*. London: Routledge. 272–285.
- Leather, J. (1999). Second language speech research: an introduction. In J. Leather (ed.) *Phonological issues in language learning*. Oxford: Blackwell. 1–58.
- Leather, J. & A. James (1996). Second language speech. In E. C. Ritchie & T. K. Bhatia (eds.) *Handbook of second language acquisition*. San Diego: Academic Press. 169–316.
- Leather, J. & A. James (eds.) (1997). *Second-language speech: structure and process*. Berlin: Mouton de Gruyter.
- Leech, G., P. Rayson & A. Wilson (2001). *Word frequencies in written and spoken English*. London: Longman.
- Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.
- Leyden, K. van (2004). *Prosodic characteristics of Orkney and Shetland dialects: an experimental approach*. PhD dissertation, Leiden University. Published, Utrecht: LOT.
- Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology* **140**, 1–55.
- Lippi-Green, R. (1997). *English with an accent: language, ideology and discrimination in the United States*. London: Routledge.
- Ludwig, J. (1982). Native-speaker judgements of second-language learners’ effort at communication: a review. *Modern Language Journal* **66**, 274–283.
- Luke, D. (2004). *Multilevel modeling*. Thousand Oaks, Ca.: Sage.
- McAllister, R. (1997). Perceptual foreign accent: L2 users’ comprehension ability. In Leather & James (1997). 119–132.

- McArthur, T. (ed.) (1992a). *The Oxford companion to the English language*. Oxford: Oxford University Press.
- McArthur, T. (1992b). Edinburgh. In McArthur (1992a). 336.
- McArthur, T. (2002). *Oxford guide to World English*. Oxford: Oxford University Press.
- McCafferty, K. (1999). (London)Derry: between Ulster and local speech – class, ethnicity and language change. In Foulkes & Docherty (1999). 246–264.
- McClure, J. D. (1994). English in Scotland. In Burchfield (1994a). 23–93.
- Macdonald, S. (2002). Pronunciation – views and practices of reluctant teachers. *Prospect: an Australian Journal of TESOL* 17. www.ncltr.mq.edu.au/prospect/17/pros17_3smac.asp. Accessed 16 May 2006.
- Mac Éinrí, P. (2001). Immigration into Ireland: trends, policy responses, outlook. Cork: Irish Centre for Migration Studies. migration.ucc.ie/irelandfirstreport.htm. Accessed 16 May 2006.
- McKay, S. L. (2002). *Teaching English as an international language: rethinking goals and approaches*. Oxford: Oxford University Press.
- MacKenzie, I. (2003). English as a Lingua Franca and European universities. *The European English Messenger* 12. 59–62.
- Major, R., S. F. Fitzmaurice, F. Bunta & C. Balasubramanian (2002). The effects of nonnative accents on listening comprehension: implications for ESL assessment. *TESOL Quarterly* 36. 173–190.
- Major, R., S. F. Fitzmaurice, F. Bunta & C. Balasubramanian (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning* 55. 37–69.
- Markham, D. (1997). *Phonetic imitation, accent and the learner*. Lund: Lund University Press.
- Marslen-Wilson, W. D. & A. Welsh (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10. 29–63.
- Mathisen, A. G. (1999). Sandwell, West Midlands: ambiguous perspectives on gender patterns and models of change. In Foulkes & Docherty (1999). 107–123.
- Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception & Psychophysics* 62. 253–265.
- Melchers, G. (2004). English spoken in Orkney and Shetland: phonology. In Schneider *et al.* (2004). 35–46.
- Melchers, G. & P. Shaw (2003). *World Englishes: an introduction*. London: Arnold.
- Mencken, H. L. (1949). *The American language: an inquiry into the development of English in the United States*. New York: Knopf.
- Mencken, H. L. (1952). *The American language: an inquiry into the development of English in the United States*. Supplement 2. New York: Knopf.
- Milroy, L. (1994). Standard English and language ideology in Britain and the United States. In Bex & Watts (1999). 173–206.

- Ministerie van OCW (2003). Kerndoelen basisvorming 1998-2003: relaties in beeld. www.minocw.nl/documenten/kerndoelen.pdf. Accessed 16 May 2006.
- Modiano, M. (1996). *A Mid-Atlantic handbook: American and British English*. Lund: Studentlitteratur.
- Morley, J. (1996). Second language speech/pronunciation: acquisition, instruction, standards, variation, and accent. In J. E. Alatis, C. A. Strahle, M. Ronkin & B. Gallenberger (eds.) *Linguistics, language acquisition, and language variation: current trends and future prospects*. Washington D.C.: Georgetown University Press. 140–159.
- Moyer, A. (1999). Ultimate attainment in L2 phonology: the critical factors of age, motivation and instruction. *Studies in Second Language Acquisition* **21**. 81–108.
- Mufwene, S. S. (2001). African-American English. In Algeo (2001). 291–324.
- Munro, M. J. & T. M. Derwing (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* **45**. 73–97.
- Nanni, D. L. (1977). Stressing words in -ative. *Linguistic Inquiry* **8**. 752–763.
- Nemser, W. (1971). Approximative systems of foreign language learners. *International Review of Applied Linguistics* **9**. 115–123. Reprinted in Richards (1974). 55–63.
- Nickel, G. (1972). Grundsätzliches zur Fehleranalyse und Fehlerbewertung. In G. Nickel (ed.) *Fehlerkunde: Beiträge zur Fehleranalyse, Fehlerbewertung und Fehlertherapie*. Berlin: Cornelsen-Verhagen & Klasing. 8–24.
- Norell, P. (1991). *Native-speaker reactions to Swedish pronunciation errors in English: pronunciation, intelligibility and attitude*. Stockholm: Almqvist & Wiksell.
- O'Connor, J. D. (1971). *Better English pronunciation*. London: Cambridge University Press.
- O'Connor, J. D. & G. Arnold (1973). *Intonation of colloquial English*. London: Longman.
- Pavlenko, A. (2002). Poststructuralist approaches to the study of social factors in second language learning and use. In V. Cook (ed.) *Portraits of the L2 user*. Clevedon: Multilingual Matters. 277–302.
- Pederson, L. (2001). Dialects. In Algeo (2001). 253–290.
- Penfield, J. (1985). *Chicano English: an ethnic contact dialect*. Amsterdam: John Benjamins.
- Penhallurick, R. (2004). Welsh English: phonology. In Schneider *et al.* (2004). 98–112.
- Piazza, L.G. (1980). French tolerance for grammatical errors made by Americans. *Modern Language Journal* **64**. 422–427.
- Pijper, J. R. de (1983). *Modelling British English intonation*. Dordrecht: Foris.
- Pisoni, D. B. & R. E. Remez (eds.) (2005). *The handbook of speech perception*. Oxford: Blackwell.

- Politzer, R. L. (1978). Errors of English speakers of German as perceived and evaluated by German natives. *Modern Language Journal* **62**. 253–259.
- Porter, D. & S. Garvin (1989). Attitudes to pronunciation in EFL. *Speak Out!* **5**. 8–15.
- Prator, C. H. (1968). The British heresy in TESL. In J. A. Fishman & J. Das Gupta (eds.) *Language problems of developing nations*. New York: Wiley. 459–476. Reprinted in Brown (1991). 11–30.
- Prator, C. H. & B. W. Robinett (1985). *Manual of American English Pronunciation*. 4th edn. New York: Holt.
- Preisler, B. (1999). Functions and forms of English in a European EFL country. In Bex & Watts (1999). 239–268.
- Preston, D. R. (ed.) (1999a). *Handbook of perceptual dialectology*. Volume 1. Amsterdam & Philadelphia: John Benjamins.
- Preston, D. R. (1999b). A language attitude approach to the perception of regional variety. In Preston (1999a). 359–373.
- Preston, D. R. (2005). How can you learn a language that isn't there? In Dziubalska-Kołaczyk & Przedlacka (2005). 37–58.
- Quené, H. & H. van den Bergh (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication* **43**. 103–121.
- Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford & T. Lewis (2000). *A user's guide to MLwiN*. London: Institute of Education. multilevel.ioe.ac.uk. Accessed 3 December 2000 (no longer accessible, May 2006). Updated version available at mlwin.com/1_10/userman.pdf. Accessed 16 May 2006.
- Richards, J. (ed.) (1974). *Error analysis: perspectives on second language acquisition*. London: Longman.
- Richards, J., J. Platt & H. Weber (1985). *Longman dictionary of applied linguistics*. London: Harlow.
- Rietveld, A. C. M. & V. J. van Heuven (2001). *Algemene fonetiek*. 2nd edn. Bussum: Coutinho.
- Rifkin, B. (1995). Error gravity in learners' spoken Russian: a preliminary study. *Modern Language Journal* **79**. 477–490.
- Rippmann, W. (1906). *The sounds of spoken English: a manual of ear training for English students*. London: Dent.
- Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association* **34**. 239–245.
- Roach, P. (2005). Representing the English model. In Dziubalska-Kołaczyk & Przedlacka (2005). 393–400.
- Romaine, S. (1980). Stylistic variation and evaluative reactions to speech: problems in the investigation of linguistic attitudes in Scotland. *Language and Speech* **23**. 213–232.
- Rooy, B. van (2004). Black South African English: phonology. In Schneider *et al.* (2004). 943–952.
- Rowling, J. K. (1999). *Harry Potter and the prisoner of Azkaban*. London: Bloomsbury.

- Ryan, E. B. (1983). Social psychological mechanisms underlying native speaker evaluations of non-native speech. *Studies in Second Language Acquisition* 5. 148–159.
- Ryan, E. B. & M. A. Carranza (1976). Attitudes toward accented English. *Atisbos: Journal of Chicano Research*. Winter 1976–77. 27–34.
- Ryan, E. B., M.A. Carranza & R. W. Moffie (1975). Mexican American reactions to accented English. In J. W. Berry & W. J. Loaner (eds.) *Applied cross-cultural psychology*. Amsterdam: Swets & Zeitlinger. 174–178.
- Ryan, E. B. & R. J. Sebastian (1980). The effects of speech style and social class background on social judgements of speakers. *British Journal of Social and Clinical Psychology* 19. 229–233.
- Ryan, E. B., R. J. Sebastian, C. Grillot & C. L. Kennedy (1980). The social utility of retaining an ethnic accent. Paper presented at the 88th Annual Meeting of the American Psychological Association, Montreal.
- Saciuk, B. (1989). Some observations on Puerto Rican phonology. *Romance Languages Annual*. tell.fll.purdue.edu/RLA-archiv/1989/Linguistics-html/Saciuk-FF.htm. Accessed 21 July 2003 (no longer accessible, May 2006).
- Sawyer, J. B. (1971). Social aspects of bilingualism in San Antonio, Texas. In H. B. Allen & G. N. Underwood (eds.) *Readings in American dialectology*. New York: Appleton Century Crofts. 375–381.
- Schairer, K. E. (1992). Native speaker reaction to non-native speech. *Modern Language Journal* 76. 309–319.
- Scheuer, S. (2005). Why native speakers are (still) relevant. In Dziubalska-Kořaczyk & Przedlacka (2005). 111–130.
- Schmied, J. (2004). East African English (Kenya, Uganda, Tanzania): phonology. In Schneider *et al.* (2004). 918–930.
- Schneider, E. W. (2004a). Global synopsis: phonetic and phonological variation in English world-wide. In Schneider *et al.* (2004). 1111–1137.
- Schneider, E. W. (2004b). Synopsis: phonological variation in the Americas and the Caribbean. In Schneider *et al.* (2004). 1075–1088.
- Schneider, E.W., K. Burridge, B. Kortmann, R. Mesthrie & C. Upton (eds.) (2004). *A handbook of varieties of English*. Volume 1: *Phonology*. Berlin: Mouton de Gruyter.
- Schuderer, A. (2002). Kontrastive Phonetik und Phonologie: suprasegmentale Merkmale. www.andreas-schuderer.de/prosodies10.pdf. Accessed 16 May 2006.
- Schwartz, G. (2005). The Lingua Franca Core and the phonetics-phonology interface. In Dziubalska-Kořaczyk & Przedlacka (2005). 177–198.
- Scovel, T. (1988). *A time to speak: a psycholinguistic inquiry into the critical period for human speech*. Cambridge, Mass.: Newbury.
- Scovel, T. (1997). Review of Singleton & Lengyel (1995). *Modern Language Journal* 81. 118–119.

- Sebastian, R. J. & E. B. Ryan (1985). Speech cues and social evaluation: markers of ethnicity, social class and age. In H. Giles (ed.) *Recent advances in language, communication and social psychology*. London: Lawrence Erlbaum. 112–138.
- Sebastian, R. J., E. B. Ryan & L. Corso (1978). Social judgements of speakers with differing degrees of accentedness. Paper presented at the 9th World Congress of Sociology, Uppsala, Sweden.
- Seidlhofer, B. (2001). Closing a conceptual gap: the case for a description of English as a Lingua Franca. *International Journal of Applied Linguistics* **11**. 133–158.
- Seidlhofer, B. (2005). Language variation and change: the case of English as a Lingua Franca. In Dziubalska-Kořaczyk & Przedlacka (2005). 59–98.
- Selinker, L. (1972). Interlanguage. In Richards (1974). 31–54.
- Selkirk, E. O. (1972). *The phrase phonology of English and French*. PhD dissertation, MIT. Published 1980, New York: Garland.
- Setter, J. & J. Jenkins (2005). Pronunciation. *Language Teaching* **38**. 1–17.
- Sheorey, R. (1986). Error perceptions of native-speaking and non-native speaking teachers of ESL. *ELT Journal* **40**. 306–312.
- Shuken, C. (1984). Highland and Island English. In Trudgill (1984b). 152–165.
- Shuken, C. (1985). Variation in Hebridean English. In M. Görlach (ed.) *Focus on Scotland*. Amsterdam & Philadelphia: John Benjamins. 145–158.
- Singleton, D. & Z. Lengyel (eds.) (1995). *The age factor in second language acquisition: a critical look at the Critical Period Hypothesis*. Clevedon: Multilingual Matters.
- Snijders, T. & R. Bosker (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage Publications.
- Solomon, D. J. (2001). Conducting web-based surveys. *Practical Assessment, Research & Evaluation* **7**. pareonline.net/getvn.asp?v=7&n=19. Accessed 16 May 2006.
- Sommers, M. S. (2005). Age-related changes in spoken word recognition. In Pisoni & Remez (2005). 469–493.
- Steenkamp, J. (1993). Etnocentrisme bij Europese consumenten. *Tijdschrift voor Management* **27**. 19–25.
- Steketee, H. (2005). The Dutch speak their languages quite well: the English language is a game for insiders. *Thema: The Netherlands, NRC Handelsblad*. 25 June 2005. 16.
- Steriade, D. (1993). Positional neutralization. Paper presented at the 24th Conference of the North East Linguistic Society, University of Massachusetts, Amherst.
- Stoddart, J., C. Upton & J. D. A. Widdowson (1999). Sheffield dialect in the 1990s: revisiting the concept of NORMs. In Foulkes & Docherty (1999). 72–89.
- Stuart-Smith, J. (2004). Scottish English: phonology. In Schneider *et al.* (2004). 47–67.

- Swacker, M. (1976). When (+native) is (–favourable). *Lektos, Special Issue*. 16–19.
- Tarone, E. (1988). *Variation in interlanguage*. London: Arnold.
- Taylor, D. M. & H. Giles (1979). At the crossroads of research into language and ethnic relations. In H. Giles & B. Saint-Jacques (eds.) *Language and ethnic relations*. Oxford: Pergamon Press. 231–242.
- Thomas, A. R. (1994). English in Wales. In Burchfield (1994a). 94–147.
- Thomas, C. K. (1947). *An introduction to the phonetics of English*. New York: Ronald Press.
- Thomas, E. R. (2004). Rural Southern white accents. In Schneider *et al.* (2004). 300–324.
- Tillery, J. & G. Bailey (2004). The urban south: phonology. In Schneider *et al.* (2004). 325–337.
- Todd, L. (1992). Anglo-Irish. In McArthur (1992a). 67–68.
- Tollfree, L. (1999). South East London English: discrete *versus* continuous modelling of consonantal reduction. In Foulkes & Docherty (1999). 163–184.
- Tottie, G. (2002). *An introduction to American English*. Oxford: Blackwell.
- Trask, R. (1996). *A dictionary of phonetics and phonology*. London: Routledge.
- Trommelen, M. (1983). *The syllable in Dutch, with special reference to diminutive formation*. Dordrecht: Foris.
- Trommelen, M. & W. Zonneveld (1973). *Inleiding in de generatieve fonologie*. Muiderberg: Coutinho.
- Trommelen, M. & W. Zonneveld (1999). Word-stress in West-Germanic: English and Dutch. In H. van der Hulst (ed.) *Word prosodic systems in the languages of Europe*. Berlin: Mouton de Gruyter. 477–514.
- Trubetzkoy, N. S. (1931). *Phonologie et géographie linguistique. Transactions du Cercle Linguistique de Prague 4*.
- Trudgill, P. (1984a). *On dialect: social and geographical perspectives*. New York & London: New York University Press.
- Trudgill, P. (ed.) (1984b). *Language in the British Isles*. Cambridge: Cambridge University Press.
- Trudgill, P. (2004). The dialect of East Anglia: phonology. In Schneider *et al.* (2004). 163–177.
- Trudgill, P. (2005a). Finding the speaker-listener equilibrium: segmental phonological models in EFL. In Dziubalska-Kołodziej & Przedlacka (2005). 213–228.
- Trudgill, P. (2005b). Native speaker segmental phonological models. In Dziubalska-Kołodziej & Przedlacka (2005). 77–98.
- Trudgill, P. & J. Hannah (2002). *International English: a guide to varieties of Standard English*. London: Arnold.
- Vaissière, J. (2005). Perception of intonation. In Pisoni & Remez (2005). 236–263.
- Veenker, T. (2003). *WWStim*. www.let.uu.nl/~theo.veenker/personal/projects/wwstim/doc/en/. Accessed 16 May 2006.

- Vogten, L. L. M. & E. Gigi (2002). *GIPOS: Graphical Interactive Processing of Speech*. Version 2.3. Eindhoven: Institute for Perception Research. www.tue.nl/ipo/hearing/gipos. Accessed 1 April 2002 (no longer accessible, May 2006).
- Wakelin, M. (1984). Rural dialects in England. In Trudgill (1984b). 70–93.
- Warner, N., A. Jongman, J. Sereno & R. Kemps (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* **32**. 251–276.
- Weiner, E. & C. Upton (2000). [hat], [hæt], and all that. *English Today* **16**. 44–45.
- Weinreich, U. (1954). Is a structural dialectology possible? *Word* **10**. 388–400.
- Weldon, T. L. (2004). Gullah: phonology. In Schneider *et al.* (2004). 393–406.
- Wells, J. C. (1970). Local accents in England and Wales. *Journal of Linguistics* **6**. 231–252.
- Wells, J. C. (1982). *Accents of English*. Cambridge: Cambridge University Press.
- Wells, J. C. (1984). English accents in England. In Trudgill (1984b). 55–69.
- Wells, J. C. (1997). Whatever happened to Received Pronunciation? In C. Medina Casado & C. Soto Palomo (eds.) *II Jornadas de Estudios Ingleses*. Jaén: Universidad de Jaén. 19–28. www.phon.ucl.ac.uk/home/wells/rphappened.htm. Accessed 16 May 2006.
- Wells, J. C. (2000). *Longman pronunciation dictionary*. 2nd edn. Harlow: Pearson Education.
- Wells, J. C. (2005). Goals in teaching. In Dziubalska-Kołaczyk & Przedlacka (2005). 101–110.
- Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly* **28**. 377–389.
- Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris.
- Windsor-Lewis, J. (1972). *A concise pronouncing dictionary of British and American English*. London: Oxford University Press.
- Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.
- Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington: Center for Applied Linguistics.
- Wolfram, W. & N. Schilling-Estes (1998). *American English: dialects and variation*. Oxford: Blackwell.
- Wrembel, M. (2005). An overview of English pronunciation teaching materials. Patterns of change: model accents, goals and priorities. In Dziubalska-Kołaczyk & Przedlacka (2005). 421–437.
- Yates, L. (2001). Teaching pronunciation in the AMEP: current practice and professional development. www.nceltr.mq.edu.au/conference2001/papers/yates_pron.ppt. Accessed 16 May 2006.

SAMENVATTING IN HET NEDERLANDS

(SUMMARY IN DUTCH)

Bij moedertaalsprekers bestaat de neiging om buitenlandse accenten over het algemeen eerder negatief dan positief te beoordelen (zie bijvoorbeeld Leather 1999). Dit kan voor niet-moedertaalsprekers een ongunstige uitwerking hebben, zowel op het persoonlijke als het zakelijke vlak. Het is voor adolescenten en volwassenen die een vreemde taal leren spreken meestal niet mogelijk (en soms niet gewenst) om hun uitspraak eenvoudigweg aan die van moedertaalsprekers aan te passen. Voor deze groepen zou het nuttig zijn te weten welke specifieke uitspraakproblemen een effectieve communicatie met moedertaalsprekers in de weg staan, zodat ze zich daarop kunnen toeleggen.

Er is onderzoek gedaan (bijvoorbeeld Johansson 1975, 1978, Dretzke 1985) naar de vraag of moedertaalsprekers bepaalde uitspraakfouten opvallender of storender vinden dan andere. Op deze manier kan worden beoordeeld of in deze fouten een rangorde is aan te brengen. Bij uitspraaktraining zou dan aan de voornaamste uitspraakproblemen in deze “foutenhierarchie” (Eng. *hierarchy of error*) de meeste aandacht gegeven kunnen worden. Een voorbeeld hiervan is de foutenhierarchie die Collins & Mees (2003b) hebben opgesteld ten behoeve van Nederlandse sprekers van het Brits Engels.

Recentelijk is de vraag gerezen of de prioritering van uitspraakproblemen wel gerelateerd moet worden aan het oordeel van moedertaalsprekers. Zo stelt Jenkins (2000) dat het Engels voornamelijk gebruikt wordt in communicatie tussen niet-moedertaalsprekers, en dat ten gevolge hiervan het oordeel van moedertaalsprekers van het Engels niet meer relevant zou zijn. Volgens deze prioriteitstelling zou men zich moeten richten op wat “internationaal” verstaanbaar is, en geen aandacht schenken aan bijvoorbeeld de Engelse **th**-klanken, die moeilijker te leren zouden zijn. In dit proefschrift wordt betoogd dat het maar de vraag is of Jenkins hiermee zowel aan de belangen als de wensen van de niet-moedertaalsprekers van het Engels tegemoet komt. Voor deze groep lijkt het eerder van belang zich te kunnen handhaven in een wereld waarin de moedertaalsprekers van het Engels ondanks hun relatief geringe aantal socio-linguïstisch de boventoon voeren.

Vanuit deze gedachte is er in dit proefschrift empirisch-experimenteel onderzoek gedaan naar de manier waarop nu juist moedertaalsprekers van het Engels reageren op een dertigtal karakteristieke Nederlandse uitspraakfouten. Deze oordelen zijn na analyse vervat in een foutenhierarchie die bij het trainen van de uitspraak gebruikt kan worden (zie hoofdstuk 5.2.2). Aan de hand hiervan kunnen Nederlandse sprekers van het Engels leren efficiënter te communiceren met diegenen die onbekend zijn met het Nederlandse Engels – niet alleen moedertaalsprekers, maar ook anderen.

Bij het onderzoek is rekening gehouden met het feit dat het Engels een groot aantal onderscheiden variëteiten kent, en dat sommige daarvan ook bij niet-moedertaalsprekers bekend zijn. Zo worden het Standaardengels uit Zuid-Engeland (ook bekend als *Received Pronunciation*) en het Standaardengels uit de Verenigde Staten (of *General American*) door veel Nederlanders bewust of onbewust als uitspraakmodel gebruikt (zie Van der Haagen 1998). Om aan deze diversiteit recht te doen, is er in dit proefschrift voor zowel *Received Pronunciation* als *General American* een foutenhiërarchie opgesteld.

Het onderzoek is niet alleen gebaseerd op de oordelen van proefpersonen die zichzelf beschreven als standaardsprekers van deze accenten; ook sprekers van andere variëteiten (zoals het Schots, Australisch, Newyorks en Canadees Engels) is gevraagd een oordeel te geven over de mate waarin Nederlandse uitspraakfouten in *Received Pronunciation* of *General American* acceptabel zijn. Dit is mogelijk gemaakt door het gebruik van een internetenquête, waaraan door ruim 500 respondenten uit de gehele Engelssprekende wereld is deelgenomen.

De verschillende oordelen van deze groepen worden in detail in dit proefschrift besproken en laten zien dat de normen van moedertaalsprekers van het Engels geen monolithisch geheel zijn, maar deels beïnvloed worden door het accent van de beoordelaars. Voorzover bekend is dit de eerste keer dat er systematisch en op grote schaal onderzoek is gedaan naar de invloed van deze taalachtergrond op het oordeel van moedertaalsprekers. Dit werd gedaan aan de hand van *multi-level* analyse, een nieuwe statistische methode die bij uitstek geschikt is voor dit soort onderzoek.

Er waren opvallende verschillen tussen de groepen respondenten in de manier waarop ze uitspraakfouten waarnamen en beoordeelden. Zo waren Noord Amerikanen aantoonbaar strenger in het beoordelen van waargenomen fouten dan andere groepen moedertaalsprekers. Daarentegen namen beoordelaars uit bijvoorbeeld de Britse eilanden en Australië aantoonbaar meer fouten waar, maar hun beoordeling daarvan was coulanter. Het is mogelijk dat Noord Amerikanen afwijkend taalgebruik sterker afwijzen dan andere groepen. Ook kan er sprake zijn van een minder indirecte houding jegens het Engels van buitenlanders dan in landen als Groot-Brittannië. Hoe dan ook, de strengere beoordeling door Noord Amerikanen past niet in het beeld dat veel Nederlanders hebben. Uit onderzoek van de auteur blijkt dat Nederlanders juist aan Britten en Ieren een grotere strengheid toeschrijven (zie 1.1). Hierdoor kan bij Nederlanders de misvatting ontstaan dat hun uitspraak van het Engels in de Verenigde Staten en Canada aan minder kritiek onderhevig zou zijn.

Ook bleek dat de ernst van uitspraakfouten in *Received Pronunciation* en *General American* soms heel anders beoordeeld werd. Dit was lang niet altijd te voorspellen door de klanksystemen van deze variëteiten met elkaar te vergelijken. In beide accenten komen **th**-klanken voor, maar voor Noord Amerikanen was het aantoonbaar bezwaarlijker als deze uitgesproken werden als /t/ of /d/. Kennelijk rust hier in de V.S. en Canada een nog groter stigma op dan in andere delen van de Engelstalige wereld. Ook andere Nederlandse realisaties van klanken die in beide accenten voorkomen (zoals een te donkere /l/ of een

niet-geaspireerde /t/) werden geheel anders beoordeeld. Op het gebruik van typisch Amerikaanse uitspraakkenmerken in een Brits accent (of vice versa) werd verschillend gereageerd, maar over het algemeen werd dit minder belangrijk gevonden. Het lijkt dus van belang dat Nederlanders (en andere niet-moedertaalsprekers van het Engels) zich rekenschap geven van de verschillen tussen Britse en Amerikaanse variëteiten van het Engels – vooral wat betreft de mate waarin bepaalde buitenlandse uitspraakfouten gestigmatiseerd worden.

Bij sommige kenmerkende Nederlandse uitspraakfouten in het Engels kan het oordeel van moedertaalsprekers beïnvloed worden door het feit dat vergelijkbare realisaties ook in verschillende nationale of regionale accenten voorkomen. Een goed voorbeeld is de Nederlandse uitspraak van *film* als “fillem”, met een extra klinker tussen de /l/ en de /m/. Iets vergelijkbaars wordt ook wel in het Iers Engels aangetroffen. Dergelijke overeenkomsten met een of meerdere accenten van het Engels werden eveneens voor negentien andere uitspraakproblemen gevonden (zie 4.2 en 4.4). Vervolgens werd het effect van deze “accent-overeenkomst” (“*accent similarity*”) onderzocht. Hieruit bleek dat als beoordelaars een vergelijkbare klankrealisatie kenden in hun eigen accent, ze Nederlandse uitspraakfouten over het algemeen coulanter beoordeelden. Dit gold echter voor een beperkt aantal gevallen; eenmaal werd zelfs het omgekeerde waargenomen. Hierin lijkt de mate waarin een bepaalde wijze van uitspreken gestigmatiseerd is een rol te spelen. Er kan dus niet zonder meer aangenomen worden dat moedertaalsprekers coulanter zijn tegenover buitenlandse uitspraakkenmerken die ook in hun eigen accent voorkomen. Dit betekent dat niet-moedertaalsprekers van het Engels beter terughoudend kunnen zijn met het gebruik van lokaal gebonden uitspraakkenmerken – zelfs in communicatie met moedertaalsprekers die deze mogelijk zelf bezigen.

Door alle groepen beoordelaars werd groot belang gehecht aan uitspraakfouten die de verstaanbaarheid verminderen. Vooral onjuiste klemtonen en foneemverwisselingen zoals (f~v, t~d, v~w, æ~e) werden als ernstige fouten gezien. (Door deze foneemverwisselingen klinkt *very* als *ferry*, *bed* als *bet*, *wine* als *vine* en *bat* als *bet*.) Dit gold echter ook voor uitspraakproblemen als het gebruik van een huig-*r* in *red*, of een extra klinker in *film*, die de verstaanbaarheid in het geheel niet in de weg staan. Kennelijk is verstaanbaarheid niet het enige criterium dat moedertaalsprekers gebruiken om buitenlandse accenten te beoordelen. Als men zich bij de prioritering van uitspraakproblemen wil richten op het oordeel van moedertaalsprekers, dan moet daarbij ook aandacht zijn voor fouten die geen probleem opleveren voor de verstaanbaarheid, maar die wel leiden tot bijvoorbeeld ergernis of vermaak. Indien het uitspraakonderwijs gebaseerd zou worden op de principes van Jenkins' (2000) *International English*, of andere studies die zich voornamelijk richten op verstaanbaarheid, dan bestaat er een gerede kans dat diegenen die ook met moedertaalsprekers willen communiceren onvoldoende op dit soort stigmatisering worden voorbereid.

De fouthiërarchieën die in dit proefschrift gepresenteerd worden zijn uitsluitend gebaseerd op het onderzoek naar de oordelen van moedertaalsprekers van het Engels dat hierboven beschreven is. Daarentegen is de *keuze* van de

onderzochte uitspraakproblemen gebaseerd op een tweede, Nederlandstalig onderzoek, dat gehouden is onder docenten Engels in het Nederlands voortgezet en hoger onderwijs, en onder studenten Engels aan een aantal Nederlandse universiteiten. Aangezien deze groepen geacht kunnen worden de meeste ervaring te hebben met Engels uitspraakonderwijs in Nederland, zijn de Nederlandse uitspraakproblemen die deze groepen het meest significant vonden opgenomen in het eerdergenoemde Engelstalige onderzoek. Hierdoor kan de selectie van de onderzochte uitspraakfouten representatief worden genoemd.

Doordat zowel de Engelstalige als de Nederlandse deelnemers een aantal dezelfde fouten beoordeeld hebben, was het in principe mogelijk deze met elkaar te vergelijken. Er waren echter ook een aantal structurele verschillen tussen de twee experimenten, wat een betrouwbare vergelijking bemoeilijkte. Als gevolg hiervan is er geen overtuigend bewijs gevonden dat de Engelstalige en Nederlandse beoordelaars consequent anders op de onderzochte uitspraakfouten reageerden: met name de studenten Engels leken met de moedertaalsprekers vaak op één lijn te zitten. Desalniettemin werd als algemene tendens gevonden dat Nederlandse beoordelaars het belang van een aantal fonemverwisselingen, maar ook van andere fouten die de verstaanbaarheid minder of niet beïnvloeden, enigszins leken te onderschatten. Bij de Nederlanders leek ook minder geneigdheid te bestaan om de ernst van een uitspraakfout af te meten aan het relatieve belang dat hieraan door bijvoorbeeld sprekers van Brits of Amerikaans Engels gehecht wordt. Het is aan te bevelen dat docenten Engels hun studenten of leerlingen ervan bewust maken dat er zowel op meta-linguïstisch als sociolinguïstisch niveau factoren zijn die de reacties van moedertaalsprekers op een Nederlandse uitspraak van het Engels beïnvloeden.

In het Nederlandse onderzoek werden ook algemene vragen gesteld over het uitspraakonderwijs in Nederland. Een vergelijking van de antwoorden van de docenten Engels in het voortgezet en hoger onderwijs (met betrekking tot hun eigen onderwijspraktijk) en de antwoorden van de studenten Engels (met betrekking tot hun middelbare-schoolervaringen met het vak Engels) bracht een aantal opmerkelijke resultaten aan het licht. In de eerste plaats wordt in het voortgezet onderwijs weinig aandacht gegeven aan uitspraak, terwijl dit volgens de exameneisen (en volgens een meerderheid van de deelnemers aan het Nederlandse onderzoek) wel degelijk onderdeel uitmaakt van de vaardigheden waarop leerlingen beoordeeld worden. Zo worden de verschillen tussen de Nederlandse en Engelse klanksystemen of tussen de variëteiten van het Engels niet of nauwelijks behandeld. In het hoger onderwijs krijgt dit beduidend meer aandacht. Ten tweede krijgen leerlingen in het voortgezet onderwijs minder oefening in het Engels dan studenten aan hogescholen en universiteiten. In een beperkt aantal gevallen werd leerlingen zelden of nooit gevraagd om bij wijze van oefening Engels te spreken. Ten slotte vond een kleine minderheid van de ondervraagde studenten dat hun leraren Engels meestal een Nederlands accent hadden.

Het is zeker niet uit te sluiten dat deze uitkomsten anders zouden zijn als ook docenten en studenten die *niet* in het onderwerp geïnteresseerd zijn hadden

deelgenomen, maar het stemt evengoed tot nadenken. Als deze resultaten de situatie van uitspraaktraining in het Nederlands voortgezet onderwijs goed weer-geven, dan lijkt deze eerder aan te sluiten bij beleid dat leerlingen bewust een Nederlandse variëteit van het Engels aan wil leren dan beleid dat leerlingen wil voorbereiden op daadwerkelijke internationale communicatie in het Engels als *vreemde taal*.

Concrete aanbevelingen ter verbetering van uitspraaktraining in zowel het voortgezet als het hoger onderwijs in Nederland worden gegeven in hoofdstuk 6.2. Wat betreft de Engelse taalvaardigheid in het hoger onderwijs is het voor diegenen die verantwoordelijk zijn voor het vaststellen, faciliteren en uitvoeren van de onderwijsprogramma's zaak op de hoogte te zijn van het geringe belang dat op middelbare scholen aan uitspraakonderwijs wordt gehecht. Als zij het noodzakelijk vinden dat afgestudeerden efficiënt kunnen communiceren met moedertaalsprekers van het Engels, dienen zij zeker te stellen dat uitspraak-training stevig verankerd blijft in taalvaardigheidsprogramma's – zowel voor anglisten en toekomstige leraren Engels als voor anderen die Engels professioneel willen gebruiken. Om te voorkomen dat beginnende studenten Engels de uitspraakfouten moeten afleren die zij onbewust in eerdere onderwijssituaties hebben aangeleerd, zou het aanbeveling verdienen als hiervoor samenwerking wordt gezocht met het voortgezet onderwijs.

CURRICULUM VITAE

Rias van den Doel was born in Amsterdam on 11 February 1965. From 1976 to 1982, he attended Johan van Oldenbarnevelt Gymnasium in Amersfoort. After receiving his *Eindexamen* (school leaving certificate) in 1982, he read English language and literature at Utrecht University, taking a special interest in Celtic languages. As part of his degree course, he spent a year at Trinity College Dublin as a Harting Scholar. In addition to following courses there in English and Anglo-Irish literature, he taught Dutch to Irish undergraduates, and for a time acquired a fine Irish accent in his English. On returning to Utrecht, he specialised in post-modern American literature, and in 1988 he received his *Doctoraal* degree (equivalent to masters) in English language and literature *cum laude*. He thereupon returned to Trinity College Dublin, to take part in a one-year MPhil programme in Anglo-Irish Literature, graduating in 1990. Subsequently, he spent a number of months at the Háskóli Íslands at Reykjavík in Iceland, where he studied Icelandic. In 1992, he was appointed to part-time lectureships in English proficiency at two universities – Utrecht and Leiden – where he also taught courses in literature and linguistics. He found time to develop an interest in amateur dramatics and took part in a number of student productions in English. In 2004, he left Leiden and took up a tenured post at Utrecht University, where he currently teaches English language and courses in applied linguistics. He now lives in Utrecht and since 2003 has been married to Maxim Brouwer.