The processing and evaluation of fluency in native and non-native speech

The research reported here was supported by Pearson Language Testing by means of a grant awarded to Nivja H. de Jong: “Oral Fluency: Production and Perception”.

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111
e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration by Benjamin A. Los

ISBN: 978-94-6093-135-2
NUR: 616

The processing and evaluation of fluency in native and non-native speech

De verwerking en beoordeling van vloeiendheid in spraak in eerste en tweede taal

(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. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
vrijdag 23 mei 2014
des middags te 12.45 uur

door

Hans Rutger Bosker

geboren 10 september 1987
te Leiderdorp

Promotor:	Prof.dr. T. J. M. Sanders
Copromotoren:	Dr. N. H. de Jong Dr. H. Quené

"Wie in zijn spreken niet struikelt,
is een volmaakt man"

Jakobus 3:2

Contents

CHAPTER 1

Introduction

1.1 The disfluent nature of speech

On 17th April 2013, Crown Prince Willem-Alexander of The Netherlands, now King, was interviewed about his future ascension to the throne. At one point, one of the interviewers posed a rather precarious question, to which the Prince struggled to reply:

> *"Nee, dit lijkt me echt iets wat niet verstandig is om hier een antwoord op te geven. Ik heb ook wel vaker in interviews gezegd: spreken is zilver, zwijgen is goud."*
>
> `Paraphrase in English:` "No, this seems to me something that is not wise to give an answer to. I have said before in interviews: speech is silver, silence is golden."

At least, that is what the official royal transcript records. The actual reply (http://www.youtube.com/watch?v=DsX4nhOwGBU) is better represented as:

> *"Nee **[uh]** dit lijkt me echt .. iets .. wat **[uh]** niet **[uh]** verstandig is om **[uh]** hier een **[uh]**, **een, een, een, een** antwoord op te geven. **[Uh...]** Ik heb ook wel vaker in interviews gezegd **[uh]**: 'spreken is zilver, zwijgen is goud'. **[Uh...]**"*

The example above illustrates the disfluent nature of spontaneous speech. Although this quoted utterance is, admittedly, a rather extreme instance of disfluent speech (in fact, the disfluent parts of the utterance make up almost half of the total recording time), disfluency is a common feature of many spoken utterances. Spontaneously produced speech contains all sorts of 'disfluencies', such as silent pauses, filled pauses (*uh*'s and *uhm*'s), corrections, repetitions (*"een, een, een, een"*), etc. As such, the disfluent character of speech reveals that planning and producing spoken utterances, however ordinary in everyday life, is not an altogether straightforward activity. Speakers have to come up with the communicative message they want to convey, they have to find the right words for their message, and finally articulate the sounds that make up those words (Levelt, 1989). Furthermore, translating thoughts into words takes place at a remarkable speed: for instance, when naming a picture, it takes only about 40 ms to retrieve a noun's initial sound once its grammatical gender has been retrieved (Van Turennout, Hagoort, & Brown, 1998). Therefore, the orchestration in real time of the cognitive tasks involved in speech planning and production has to take place with millisecond precision. Taking into account the time pressure under which speech production takes place, it is not surprising that speakers sometimes fail to produce fluent contributions to a conversation.

1.2 The perception of fluency

The present dissertation studies the disfluent character of speech from the perspective of *the listener*. It is commonly assumed that speech comprehension is hindered by the disfluent nature of spontaneous speech. For instance, disfluencies are often absent in transcripts and they are commonly taken out from radio interviews prior to broadcasting (so-called de-uhm'ing). Professional speakers strive hard to refrain from producing filled pauses such as *uh*'s and *uhm*'s; for instance, in all of the recorded inaugural speeches by US presidents between 1940 and 1996, there is not a single *uh* or *uhm* (Clark & Fox Tree, 2002). The assumption that listeners are hindered by disfluency is also found in the language learning community. Many language learners strive hard to speak a language fluently thus hoping to improve their comprehensibility. Research provides evidence that disfluent non-native speech negatively affects the impression that listeners have of the non-native speaker. These studies represent the evaluative approach to the study of fluency, as explained in the following paragraphs. Studies adopting this approach study fluency as a global property of the spoken discourse as a whole and have primarily focused on non-native speech.

However, another approach to the perception of fluency may be discerned. In contrast to the negative effects of disfluency on listeners' impressions, there are also indications in the literature that disfluencies may in fact help, rather than hinder, the listener in speech comprehension. The field of psycholinguistics has provided evidence for these beneficial effects of disfluencies. These studies will be referred to as the cognitive approach to fluency. Scholars adopting this approach study disfluency as a local property of a particular utterance and have primarily focused on native speech.

This dissertation combines the evaluative and the cognitive approach to come to a better understanding of the perception of native and non-native fluency. Below, both approaches will be introduced and it will be explained how the present dissertation combines both approaches.

1.3 An evaluative approach to fluency

The evaluative approach to fluency has as its goal to find valid and reliable ways of assessing speakers' language proficiency, and is concerned with fluency as a component of speaking proficiency. It views fluency as a global property of the spoken discourse as a whole. This approach primarily focuses on the evaluation of non-native speakers' speaking proficiency. This approach is taken up, for instance, in language testing practice, where human raters frequently assess non-native speakers' fluency levels (examples of such tests are TOEFL iBT, IELTS, PTE Academic). One of the central issues for the evaluative approach to fluency is to define what is to be understood by 'fluent' speech. Fillmore (1979) distinguished four different dimensions of fluent speech: (1) rapid, connected speech (e.g., a sports announcer); (2) dense, coherent speech (e.g., an eloquent scholar); (3) appropriate, relevant speech (e.g., a professional interviewer); and (4) creative, aesthetic speech (e.g., a poet or professional writer). Fillmore's distinctions do not only focus on the form of the speech, but also on its content (e.g., its relevance or coherence). In order to discriminate between, on the one hand, the form and, on the other hand, the content of speech, Lennon (1990) coined definitions of two senses of fluency. Fluency *in the broad sense* is often used as a synonym for global language ability, for instance in such statements as "He is fluent in four languages". It functions as a cover term for overall speaking proficiency (Chambers, 1997) and may refer to anything from error-free grammar to large vocabulary size or near-native pronunciation skills. In contrast, fluency *in a narrow sense* is a component of speaking proficiency. This sense is often encountered in oral examinations: apart from grammar and vocabulary, the flow and smoothness of the speech is also assessed. It is this narrow sense of fluency that this dissertation is concerned with.

1.3.1 Fluency in the narrow sense

Unfortunately, there is a myriad of definitions of fluency in the narrow sense. It has been defined as an "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and smoothly" (Lennon, 1990, p. 391). In this definition, fluency is taken, primarily, as a subjective "impression on the listener's part" rather than being a property of the speech itself. In a later publication, Lennon introduced another working definition of fluency, namely fluency as "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (Lennon, 2000, p. 26). Arguing from this definition, fluency is identified as an automatic procedural skill of the speaker (cf. Schmidt, 1992, p. 358). The interpretation of fluency by Lennon (2000) appears to pertain to both performance characteristics ("rapid, smooth") as well as linguistic competence ("accurate"). Later descriptions of fluency primarily focus on fluency as a performance feature of speech production. For instance, Housen and Kuiken (2009) state that "fluency is primarily related to learners' control over their linguistic L2 knowledge, as reflected in the speed and ease with which they access relevant L2 information to communicate meanings in real time" (p. 462). Here, fluency is again associated with cognitive speech production processes, such as linguistic control and access. Finally, Skehan (2009) has provided an interpretation of fluency that is primarily concerned with the form of the utterance, namely "the capacity to produce speech at normal rate and without interruption" (Skehan, 2009, p. 510). In this view, fluency is an acoustic phenomenon that can be measured as a property of the spoken utterance itself.

This multitude of definitions, meant to delineate the concept of fluency, rather reveals the complex and multidimensional nature of fluency. However, it is possible to discern several patterns. Some studies place the emphasis on the efficiency of the cognitive processes responsible for (dis)fluency. Others focus on the acoustic consequences of these cognitive processes for the spoken utterance. Again others stress the effect that (dis)fluent speech may have on the listener. Segalowitz (2010) tried to distinguish the different interpretations of fluency by means of one framework, proposing a cognitive science approach to fluency.

1.3.2 A fluency framework

In the fluency framework of Segalowitz (2010) the insights from various scientific disciplines are brought together (e.g., behavioral and brain sciences, social sciences, formal disciplines, philosophy of mind). In his monograph, Segalowitz argues that sociolinguistic (social context), psycholinguistic (the neurocognitive system of speech production) and psychological (e.g., motivational) factors in-

terlinked in a dynamical system all contribute to a speaker's fluency level. He describes a framework for thinking about fluency in which three interpretations of fluency are distinguished, namely *cognitive fluency* – "the efficiency of operation of the underlying processes responsible for the production of utterances"; *utterance fluency* – "the features of utterances that reflect the speaker's cognitive fluency" which can be acoustically measured; and *perceived fluency* – "the inferences listeners make about speakers' cognitive fluency based on their perceptions of the utterance fluency" (Segalowitz, 2010, p. 165). Adopting the fluency framework of Segalowitz (2010), we will summarize the literature on (the relationships between) cognitive, utterance, and perceived fluency.

Cognitive fluency A speaker's cognitive fluency is defined as the operation efficiency of speech planning, assembly, integration and execution (Segalowitz, 2010). Segalowitz adopts the speech production model of Levelt (1989). This is the most influential model of speech planning and production and it is comprised of three main phases, namely *conceptualization*, *formulation*, and *articulation* (Levelt, 1989; Levelt, Roelofs, & Meyer, 1999). A speaker wanting to convey a communicative message starts planning his/her utterance through conceptual preparation. He will plan what to say, which language to use, integrating knowledge about the sociopragmatic aspects of the conversational situation. Furthermore, the speaker comes up with a preverbal message: a conceptual structure that can be implemented in words. This preverbal message reflects how the speaker construes the communicative event, taking into account the position of the speaker and listener, the emphasis the speaker wishes to convey, etc. During the phase of formulation, the preverbal message is encoded in a grammatical form, resulting in a surface structure of the to-be-produced utterance. The surface structure forms the input to morpho-phonological encoding (choosing the right words with the correct word forms) and phonetic encoding (building an appropriate phonetic gestural score). Finally, in the articulatory phase, the articulatory plan is used to produce the required phonetic events. Segalowitz (2010) argues that the different stages in speech production form potential loci of processing difficulties which may give rise to disfluency. He terms these critical points in speech production 'fluency vulnerability points' (Segalowitz, 2010, Figure 1.2). For instance, disfluency can originate from trouble in finding out what to say, in choosing the right words, and/or in generating a phonetic plan. Therefore, the efficiency of the cognitive processes involved in speech planning and production define the fluency of the utterance.

The model of speech planning and production by Levelt (1989) is a blueprint of the monolingual speaker. This model has been adapted by De Bot (1992) to a model for the bilingual speaker (cf. Kormos, 2006). Producing speech in a second language (L2) resembles the processes involved in speaking one's

native language (L1), because speech production in an L2 also involves the conceptualization of the message, formulation of the words and articulation of the sounds. However, De Bot (1992) identified several points in Levelt's model that have particular relevance for L2 speech. De Bot (1992) assumes that some of the processes involved in conceptualization are non-language specific, such as the process of *macroplanning* which involves the elaboration of the communicative intention at the level of conceptual and propositional message content. It is assumed that encyclopedic and social knowledge is not organized in language specific terms and, as a consequence, Segalowitz (2010) argues that no L2-specific fluency issues can arise at this stage in speech production. In other words, native and non-native speakers are expected to encounter the same sorts of difficulties in macro-planning in conceptualization. In contrast, the construction of the preverbal message through *microplanning* - assigning a particular information structure to the macroplan - is thought to be language specific. Moreover, the representations in other stages of the model are presumed to be language specific, such as L1 vs. L2 lemma's, morpho-phonological codes, gestural scores, etc.

Two possible sources are thought to be responsible for the L2-specific difficulties in speech formulation and articulation. First, an L2 speaker may experience trouble because of incomplete *knowledge* of the L2 (e.g., a small L2 vocabulary, unknown grammatical rules, etc.). Second, the L2 speaker could also have insufficient *skills* with which L2 knowledge is used (e.g., lexical access, speed of articulation, etc.). Both insufficient declarative (knowledge) and procedural (skill) mastery of the L2 can lead to a decrease in specifically L2 cognitive fluency (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012a; Paradis, 2004; Towell, Hawkins, & Bazergui, 1996). Thus, it is at the stages of formulation and articulation that non-native speech is all the more vulnerable to disfluency.

Utterance fluency Utterance fluency, the acoustic manifestation of (dis)-fluency, may be considered as the most tangible interpretation of fluency. Researchers have identified a great number of phonetic measurements that may be associated with fluency, such as speech rate, mean length of runs, number of corrections or repetitions per minute, number of silent or filled pauses per minute, mean length of pauses, etc. (cf. Table 1.1 from Segalowitz, 2010, p. 6). There is also a large diversity in the way researchers calculate specific measures. In order to counter the abundance and diversity of acoustic measures, measures of utterance fluency have been clustered into three acoustic dimensions (Skehan, 2003, 2009; Tavakoli & Skehan, 2005): *breakdown fluency* concerns the extent to which a continuous speech signal is interrupted by (silent and filled) pauses; *speed fluency* has been characterized as the rate of speech delivery; and

repair fluency relates to the corrections and repetitions present in the speech signal. Nevertheless, this classification of particular acoustic fluency measures as components of either speed, breakdown or repair fluency is by no means straight-forward. For instance, the measure speech rate - calculated as the total number of syllables in a speech excerpt divided by the total recording time (including silent pauses) - is dependent on both the speaker's speed of articulation and the total number and duration of pauses. As such, the measure speech rate should be categorized as both a measure of the dimension of speed fluency and the dimension of breakdown fluency.

Perceived fluency The third and final interpretation of fluency is perceived fluency – "the inferences listeners make about speakers' cognitive fluency based on their perceptions of the utterance fluency" (Segalowitz, 2010, p. 165). Perceived fluency is most commonly assessed by means of subjective judgments, usually involving ratings on Equal Appearing Interval Scales (EAIS; Thurstone, 1928). For one of the few examples using Magnitude Estimation, see McColl and Fucci (2006). Most studies into perceived fluency have investigated the relationship between perceived fluency (subjective judgments) and utterance fluency (temporal speech measures) in order to assess the relative contributions of different speech characteristics to fluency perception. These studies indicate that temporal measures alone can account for a large amount of variance in perceived fluency ratings. Rossiter (2009) reports a correlation of $r = 0.84$ between subjective fluency ratings and pruned number of syllables per second. She also compared ratings from untrained and expert fluency raters and did not find a statistically significant difference between the two groups. Derwing, Rossiter, Munro, and Thomson (2004) used novice raters for obtaining perceived fluency judgments. These raters listened to speech materials of 20 beginner Mandarin-speaking learners of English. Derwing et al. (2004) found that pausing and pruned syllables per second together accounted for 69% of the variance of their fluency ratings. Kormos and Dénes (2004) related acoustic measurements from L2 Hungarian speakers to fluency ratings by native and non-native teachers. They report on a correlation of $r = 0.87$ between the measure speech rate and subjective fluency ratings. Cucchiarini, Strik, and Boves (2002) had teachers rate spontaneous speech materials obtained from non-native speakers of Dutch. They found a correlation of $r = 0.65$ between the mean length of runs and the perceived fluency of spontaneous speech.

These studies suggest that temporal factors are major contributors to fluency judgments. However, many researchers have raised the question whether non-temporal factors, such as grammatical accuracy, vocabulary use, or foreign accent, should also be considered as influencing fluency judgments (Freed, 1995; Lennon, 1990). Rossiter (2009) notes that subjective ratings of fluency, in her

study, were influenced by non-temporal factors as well (on the basis of qualitative analysis of rater comments). The most important factor in this respect was learners' L2 pronunciation. More recently, a quantitative study by Pinget, Bosker, Quené, and De Jong (in press) has tackled the relationship between perceived fluency and perceived accent. This study suggests that raters can keep the concept of fluency well apart from perceived foreign accent. Fluency ratings and accent ratings of the same speech samples were found to correlate only weakly ($r = -0.25$) and, moreover, acoustic measures of accent did not add any explanatory power to a statistical model of perceived fluency. This suggests that, although the contribution of non-temporal factors to perceived fluency should not be ignored, these non-temporal factors only play a minor role.

The diversity in both methodology and results of the studies into perceived fluency hinders interpretation and practical application. First of all, most studies report correlations between utterance fluency measures and perceived fluency ratings. However, empirically observed co-occurrence is a necessary but not a sufficient condition for causality, as correlation does not necessarily imply causation. Secondly, depending on the amount of detail in speech annotations, the number of available acoustic predictors of speaking fluency may grow very large. This raises the question which measures are relevant and which measures are irrelevant factors in fluency perception. This question is very difficult to answer due to the large intercollinearity of acoustic measures, which confounds the different measures. For instance, the aforementioned measure of speech rate (number of syllables divided by total time including silences), and the mean duration of silent pauses both depend on the duration of silent pauses in the speech signal, and as a result, these two measures are interrelated. If a study would find these two measures to be strongly related to fluency ratings, the relative contribution of each measure to perceived fluency remains unclear. In order to understand what raters really listen to when evaluating oral fluency, correlations among acoustic measures should also be taken into account. Unfortunately, correlations between fluency measures are often not reported in the literature, even though the degree of intercollinearity of measures may distinguish orthogonal from confounded measures. Thus, the evaluative approach to fluency calls for studies of fluency perception that use measures of utterance fluency with low intercollinearity that can distinguish between the different acoustic dimensions of utterance fluency (i.e., breakdown, speed, and repair fluency; Skehan, 2003, 2009; Tavakoli & Skehan, 2005).

1.3.3 L1 fluency

The literature review above reveals that the evaluative approach to fluency has primarily focused on the level of fluency of non-native speakers. This is most

likely due to the grounding of this approach in language testing practice, where the fluency level of non-native speakers is assessed. It is a common assumption that native speakers supposedly are perceived as fluent by default (cf. Davies, 2003; Raupach, 1983; Riggenbach, 1991). Nevertheless, native speakers clearly do not only produce fluent speech (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Raupach, 1983; Riggenbach, 1991). This raises the question what it is that distinguishes native fluency from non-native fluency.

Within the evaluative approach to fluency, there have been relatively few studies that have included native speech in their fluency research. Some use native fluency levels as controls in studies of L2 fluency perception or native speech samples are used as 'anchor stimuli' that are thought to keep the reference standard stable (e.g., Cucchiarini, Strik, & Boves, 2000). From this work, we gather that natives are consistently rated higher than non-natives (Cucchiarini et al., 2000) and that they also produce fewer disfluencies than non-natives do (Cucchiarini et al., 2000). However, from these studies we cannot gather whether the distinction between native and non-native fluency is gradient (natives produce fewer disfluencies than non-natives) or categorical (native disfluencies are weighed differently from non-native disfluencies). Hulstijn (2011) discusses the difference between native and non-native speech, suggesting that the distinction may be a gradient rather than a categorical one. Such a conclusion would carry implications for language testing practice. The fluency level of non-native speakers is regularly assessed in language tests as a component of overall L2 speaking proficiency. These tests typically evaluate L2 fluency according to native standards, revealing a hidden assumption that native speakers are a homogenous group and that nativelike performance is the final goal for non-native speakers. But if the variation in native disfluency production carries consequences for how fluent they are perceived, assessment on grounds of native norms is questionable.

1.4 A cognitive approach to fluency

Fortunately, we are not altogether uninformed when it comes to native fluency. There is a considerable body of psycholinguistic research investigating fluency in native speech, adopting a cognitive approach. The goal of the cognitive approach is to determine the cognitive factors that are responsible for disfluency (in production), and to understand how disfluent speech affects the cognitive processes of the listener (in perception), such as prediction, memory, and attention. Ever since the 1950's (e.g., Goldman-Eisler, 1958a, 1958b) scholars have investigated fluency characteristics (e.g., silent pauses, errors, repairs, etc.). Instead of focusing on the (dis)fluent character of the spoken discourse as a whole, the cognitive approach to fluency targets local phenomena, namely disfluencies.

Disfluencies have been defined as "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (Fox Tree, 1995), such as silent pauses, filled pauses (e.g., *uh* and *uhm*), corrections, repetitions, etc. The production literature has revealed that disfluencies are common in spontaneous speech: it is estimated that six in every hundred words are affected by disfluency (Bortfeld et al., 2001; Fox Tree, 1995). Therefore, researchers have traditionally argued that the disfluent character of spontaneous speech poses a challenge to the cognitive mechanisms involved in speech perception (Martin & Strange, 1968). Disfluencies were assumed to pose a continuation problem for listeners (Levelt, 1989), who were thought to be required to edit out disfluencies in order to process the remaining linguistic input. Thus, disfluencies would uniformly present obstacles to comprehension and would need to be excluded in order to study speech comprehension in its 'purest' form (cf. Brennan & Schober, 2001). However, more recent research in the field of speech perception seems to converge on the conclusion that disfluencies may help the listener in comprehension. The potentially beneficial effect of disfluency on speech comprehension, seems to originate from certain regularities in the production of disfluencies. First the work on disfluency production will be introduced before moving on to disfluency effects on speech perception.

1.4.1 Producing disfluencies

There are several factors influencing disfluency production. Some speaker characteristics have been found to affect the production of disfluencies, such as age and gender, but also the speaker's conversational role and conversational partner (Bortfeld et al., 2001). Furthermore, disfluencies have a higher probability of occurrence before linguistic content with higher cognitive load. This causes disfluencies in spontaneous speech to follow a non-arbitrary distribution: they tend to occur before longer utterances (Oviatt, 1995; Shriberg, 1996), before unpredictable lexical items (Beattie & Butterworth, 1979), before low-frequency color names (Levelt, 1983), open-class words (Maclay & Osgood, 1959), names of low-codability images (Hartsuiker & Notebaert, 2010), or at major discourse boundaries (Swerts, 1998). Also talking about an unfamiliar topic (Bortfeld et al., 2001; Merlo & Mansur, 2004) or at a higher pace (Oomen & Postma, 2001) increases the likelihood of disfluencies.

Another factor influencing disfluency production is context. It has been observed that there is a higher probability of disfluency when talking in dialogue vs. monologue and to humans vs. computers (Oviatt, 1995). In contexts where there are multiple reference options to choose from, such as in case of low contextual probability (Beattie & Butterworth, 1979) or multiple reference options (Schnadt & Corley, 2006), disfluencies are also more likely to occur. It has even been observed that lectures in the humanities are typically more

disfluent than those in the exact sciences due to the linguistically more complex nature of the humanities (Schachter, Christenfeld, Ravina, & Bilous, 1991; Schachter, Rauscher, Christenfeld, & Crone, 1994).

Judging from the reviewed literature, we find that cognitive load and context are responsible for certain regularities in the distribution of disfluencies. The fluency framework described in Segalowitz (2010) may account for the non-arbitrary distribution of disfluencies. In this framework, disfluency typically originates from difficulty at the different stages in speech production. It is at loci of relatively high cognitive load that disfluencies occur, thus explaining the non-arbitrary distribution of disfluencies in native speech. This observation is critical for our understanding of the perception of disfluencies.

1.4.2 Perceiving disfluencies

Research on speech comprehension has revealed that listeners are sensitive to the regularities in disfluency production. Listeners may use the increased likelihood of speakers to be disfluent before linguistic content with higher cognitive load as a cue to guide their expectations. For instance, the higher probability of disfluencies occurring before more complex syntactic phrases may help comprehenders to avoid erroneous syntactic parsing (Brennan & Schober, 2001; Fox Tree, 2001). Disfluencies may also aid listeners in attenuating context-driven expectations about upcoming words (Corley, MacGregor, & Donaldson, 2007; MacGregor, Corley, & Donaldson, 2010) or may improve recognition memory (Collard, Corley, MacGregor, & Donaldson, 2008; Corley et al., 2007; MacGregor et al., 2010). Finally, disfluencies have also been found to guide prediction (Arnold, Fagnano, & Tanenhaus, 2003; Arnold, Hudson Kam, & Tanenhaus, 2007; Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Barr & Seyfeddinipur, 2010; Kidd, White, & Aslin, 2011a, 2011b; Watanabe, Hirose, Den, & Minematsu, 2008). In the eye-tracking experiments of Arnold et al. (2004) using the Visual World Paradigm (Huettig, Rommers, & Meyer, 2011; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), participants were presented with discourse-old (i.e., previously mentioned) and discourse-new (i.e., not previously mentioned) referents. When presented with a disfluent utterance (e.g., 'Now move thee uh candle...'), participants' eye fixations showed, prior to target onset, a preference for the discourse-new referent (Arnold et al., 2003, 2004; Barr & Seyfeddinipur, 2010). This suggests that listeners use the increased likelihood of speakers to be disfluent while referring to new as compared to given information (Arnold, Wasow, Losongco, & Ginstrom, 2000) as a cue to the information structure of the utterance. The preference for a particular referent on the basis of the presence of a disfluency has been termed the *disfluency bias*. This bias has been shown to apply to prediction of discourse-new, but also of unknown referents (Arnold et al., 2007; Kidd et al., 2011a,

2011b; Watanabe et al., 2008). Upon presentation of a disfluent sentence such as 'Click on thee uh red [target]', there were, prior to target onset, more looks to an unknown object (an unidentifiable symbol) than a known object (e.g., an ice-cream cone), as compared to the same instruction in the fluent condition (Arnold et al., 2007).

Follow-up experiments in Arnold et al. (2007) and Barr and Seyfeddinipur (2010) targeted the cognitive processes responsible for the disfluency bias. They found that the predictive mechanisms of the listener take the speaker's perspective into account. When participants were told that the speaker they were about to hear, suffered from object agnosia - a medical condition involving difficulty recognizing simple objects - the disfluency bias for unknown objects was found to disappear (Arnold et al., 2007, Experiment 2). This suggests listeners actively make rapid inferences about the source of the disfluency. In doing so, listeners take the speaker's cognitive state into account, which modulates the extent to which disfluency guides prediction.

Because listeners use disfluency as a cue to upcoming dispreferred or complex information, they also integrate unpredictable target words more easily into a disfluent context than a fluent context. Corley et al. (2007) presented participants in their ERP experiment with highly constrained sentences such as 'She hated the CD, but then she's never liked my taste in [*uh*] music/clothes'. The ERP data revealed a classical N400 effect for the unpredictable (e.g., 'clothes') relative to the predictable condition (e.g., 'music'), indicating difficulty in integrating the unpredictable word into the sentence context. However, when the target word was preceded by a disfluency (e.g., *uh*), the N400 effect was strongly reduced (also applies to silent pauses; MacGregor et al., 2010). Apparently, the unpredictability of the target word was attenuated based on the presence of the disfluency. This suggests that listeners are aware of the increased likelihood of unpredictable content following a disfluency and that this awareness reduced the listeners' surprise upon encountering the unpredictable target.

These eye-tracking and ERP experiments demonstrate short-term effects of disfluency: hesitations affect the way in which listeners process spoken language in real time. Disfluencies have also been found to have longer-term effects regarding the retention of words immediately following disfluencies. For instance, after the ERP experiments reported in Corley et al. (2007) and MacGregor et al. (2010) participants took part in a surprise memory test. Participants were presented with written words and indicated whether they thought this word was 'old' (had occurred in the ERP experiment) or 'new' (had *not* occurred in the ERP experiment). Half of the old words had been presented in a fluent context and the other half in a disfluent context (i.e., following *uh*). It was observed that participants were more accurate in recalling old words when this word had been preceded by a disfluency (relative to the fluent condition). The authors

argue that the change in the N400, indicating a difference between the processing of fluent vs. disfluent speech, resulted in changes to the representation of the message. This idea is also supported by data from the Change Detection Paradigm (CDP) in Collard (2009, Experiments 2-6). In this paradigm, participants listen to speech passages which they try to remember. After listening to the speech, a textual representation of the passage is presented which either matches the spoken passage or contains a one word substitution. Participants, then, indicate whether they detect a change in the text or not. In the CDP reported in Collard (2009), the to-be-substituted words (i.e., target words) in the spoken passages were either presented in a fluent context or a disfluent context, with a filled pause (e.g., *uh*) preceding the target word. Collard (2009) found that listeners were more accurate at detecting a change in a CDP when the target word had been encountered in the context of a preceding filled pause hesitation.

1.4.3 Disfluencies triggering attention

The literature shows that disfluency may have short-term (prediction) and long-term effects (memory). But what is the relationship between these two types of effects? There is some evidence in the literature that suggests that listeners, upon encountering a disfluency, raise their attention to the incoming speech signal (Collard, 2009; Collard et al., 2008). Considering that disfluency introduces novel, dispreferred or more complex information, listeners may benefit from these expectations by raising their attention as a precautionary measure to ensure timely comprehension of the unexpected information. If disfluency triggers heightened attention, this might account for the beneficial effect of disfluency on the recognition of words immediately following the disfluency.

Indirect support for an attentional account of disfluency effects may be found in lower reaction times (RTs) for recognition of words immediately following a disfluency (Corley & Hartsuiker, 2011; Fox Tree, 2001), and faster responses to disfluent instructions (Brennan & Schober, 2001). Direct evidence that disfluencies affect attention has been provided by Collard et al. (2008). Participants in this study listened to sentences that sometimes contained a sentence-final target word that had been acoustically compressed, thus perceptually deviating from the rest of the sentence. This acoustic deviance induced ERP components associated with attention (mismatch negativity [MMN] and P300). However, when the deviant target word was preceded by a disfluency, the P300 effect was strongly reduced. This suggests that listeners were not required to reorient their attention to deviant words in disfluent cases. Moreover, a surprise memory test established, once again, a beneficial effect of disfluency on the recognition of previously heard words. Taken together, these results manifest the central role of attention in accounting for how disfluency is pro-

cessed: due to their non-arbitrary distribution, disfluencies cue more complex information and, therefore, capture listeners' attention. Heightened attention, then, affects the recognition and retention of words following the disfluency.

1.4.4 L2 disfluencies

The empirical evidence introduced above shows that listeners are aware of the regularities in disfluency production: when presented with disfluent speech, listeners anticipate reference to a more cognitively demanding concept. These psycholinguistic studies have, however, focused exclusively on disfluencies produced by native speakers. It is, as yet, unknown how native listeners process the disfluencies produced by non-native speakers. In speech planning, non-native speakers may experience high cognitive load where a native speaker would not. As argued above, this may be due to incomplete *knowledge* of the L2 (e.g., a small L2 vocabulary, unknown grammatical rules, etc.) or insufficient *skills* with which L2 knowledge is used (e.g., lexical access, speed of articulation, etc.). As a result, the distribution of disfluencies in non-native speech may be argued to be more irregular than the disfluency distribution in native speech (that is, from the native listener's point of view). Non-native disfluencies may be perceived by native listeners as being 'less informative' of the kind of word to follow than native disfluencies are, and as such have a differential effect on listeners' predictive strategies. If listeners take the speaker's perspective and knowledge into account in speech comprehension (see Arnold et al., 2007; Barr & Seyfeddinipur, 2010; Hanulíková, Van Alphen, Van Goch, & Weber, 2012), we may find that non-native disfluencies do not affect L1 cognitive processes in the same way as native disfluencies.

Previous psycholinguistic work on the effect of native disfluencies on prediction have studied listeners' attribution of disfluencies to speaker difficulty in *conceptualization*. More specifically, listeners attribute disfluencies to speech production difficulties with (i) recognizing unknown objects (e.g., 'I think the speaker is disfluent because she has trouble recognizing the target object'; Arnold et al., 2007; Watanabe et al., 2008) or with (ii) pragmatic status (e.g., 'I think the speaker is disfluent because she has trouble conceptualizing a discourse-new referent'; Arnold et al., 2004; Barr & Seyfeddinipur, 2010). However, it is not at this stage of speech planning that non-native speakers diverge from native speakers (De Bot, 1992). Rather, one expects to find sources of L2-specific disfluency at the stage of *formulation*. For instance, L2-specific disfluencies may arise as a consequence of the non-native speaker encountering more difficulty in accessing L2 lemma's (relative to a native speaker) during the creation of the surface structure (i.e., lexical retrieval). Therefore, if an empirical study is to find a difference in the perception of native and non-native disfluencies, one should target listeners' attributions of disfluency to difficulty

in formulation (e.g., lexical retrieval). Such a study is not only valuable for our understanding of the perception of non-native disfluencies. Attribution of native disfluencies to difficulty in other stages than conceptualization has, so far, not been reported. The study described above could shed light on the flexibility with which listeners attribute the presence of disfluency to other stages in speech production, such as formulation.

1.5 Combining evaluative and cognitive approaches

The studies in this dissertation combine the evaluative approach (Chapters 2-3) and the cognitive approach (Chapters 4-5). The chapters adopting the evaluative approach study the listener's subjective impression of the fluency level of both native and non-native speakers. The chapters adopting the cognitive approach study the listener's cognitive processes involved in comprehension of both native and non-native speech. In this fashion, it will be possible to compare how fluency characteristics in native and non-native speech contribute to the assessment of fluency, as well as to such cognitive processes as prediction, memory, and attention. This dissertation aims to resolve the apparent contradiction in the literature between, on the one hand, the *negative* effects of non-native disfluencies on subjective fluency ratings, and, on the other hand, the *positive* effects of native disfluencies on speech perception. Therefore, the following research question is formulated:

> Main RQ: How do fluency characteristics affect the perception of native and non-native speech?

1.5.1 Chapter 2

Chapter 2 aims to identify the acoustic factors that make speech sound fluent. The methodological diversity in studies relating utterance fluency to perceived fluency hinders our understanding of the acoustic correlates of perceived fluency and precludes generalization of research findings. Chapter 2 describes several experiments that address this diversity. The first experiment reported in Chapter 2 relates perceived fluency judgments to utterance fluency measures, similar to the work on perceived fluency introduced above (e.g., Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009). However, some studies have used large numbers of acoustic predictors without accounting for the potential intercollinearity of these predictors, thus threatening the validity of results. Therefore, the first experiment of Chapter 2 used only a limited set of acoustic measures, which had been particularly

selected for their low intercollinearity. Furthermore, these acoustic predictors were clustered into the three acoustic dimensions of utterance fluency, namely speed fluency, breakdown fluency, and repair fluency. Through a comparison of the independent contributions of the three acoustic dimensions to perceived fluency, it is possible to formulate an answer to the first research question of Chapter 2:

> RQ 1A: What are the independent contributions of the three fluency dimensions of utterance fluency (breakdown, speed, and repair fluency) to perceived fluency?

Also, three other experiments sought to account for the results from the first experiment by investigating listeners' perceptual sensitivity. These experiments assessed whether listeners' perceptual sensitivity to acoustic pause, speed and repair phenomena may explain their relative contributions to perceived fluency. For this, a second research question was formulated:

> RQ 1B: How well can listeners evaluate the pause, speed, and repair characteristics in speech?

The answer to RQ 1B may help interpret the findings about RQ 1A. If, for instance, pause measures can be found to be strongly related to perceived fluency ratings, the question can be posed whether this might be due to the fact that listeners are in general more sensitive to pause phenomena. If this is corroborated by the data, then perception 'paves the way' for assessment: the way we perceive speech directly influences our subjective impression of that speech. If, in contrast, there is an asymmetry between speech features that contribute to fluency perception and the features in speech listeners are most sensitive to (e.g., pause characteristics are well perceived but contribute only little to fluency perception), then perceptual sensitivity is not the only factor determining fluency perception. Listeners, in this scenario, would first perceive the acoustic characteristics of a speaker's speech but then subsequently also weigh their importance for fluency assessment. Both hypotheses would carry implications for language testing practice and for language learners. A hierarchy of the relative relevance of the different acoustic fluency dimensions for fluency perception may prove useful, for instance, for automatic fluency assessment. Also, it may potentially help language learners to prioritize improvements in one acoustic dimension over another.

1.5.2 Chapter 3

Chapter 3 builds on Chapter 2 by comparing the way listeners assess the fluency level of native and non-native speakers. Disfluencies occur both in native and

non-native speech, but most of the literature on perceived fluency has targeted the assessment of non-native fluency. By including native speech materials in the rating experiments in Chapter 3, it is possible to address the following research question:

> RQ 2: Do listeners evaluate fluency characteristics in the same way in native and non-native speech?

Native and non-native speech differ in a large range of linguistic aspects (vocabulary, grammar, pronunciation, etc.). Consequently, a valid comparison between native and non-native fluency is only viable if the native and non-native speech can be matched for the acoustic dimension under investigation. Therefore, the experiments in Chapter 3 involve phonetic manipulations in native and non-native speech. This allows for maximal control over the speech stimuli. If different fluency ratings are given to two items differing in a single manipulated phonetic property, then the perceptual difference may be reliably attributed to the minimal acoustic difference between the items. The first experiment of Chapter 3 investigates the contribution of pause incidence and pause duration to the perception of fluency in native and non-native speech by systematically altering silent pause durations. The second experiment manipulates the speed of the native and non-native speech. Chapter 3 aims to reveal how listeners judge the fluency level of native speakers and will allow for a comparison between native and non-native fluency perception. The results from Chapter 3 may potentially reveal variation in the perceived fluency of native speakers, thus complicating L2 proficiency assessment on grounds of idealized, fixed native norms.

1.5.3 Chapter 4

Chapter 4 will build on the results from Chapter 3 by adopting a cognitive approach to the perception of disfluencies. Where Chapter 3 investigates how native and non-native disfluencies affect listeners' subjective impressions of the speaker, Chapter 4 evaluates the effect that these native and non-native disfluencies have on prediction. The psycholinguistic literature on disfluencies has investigated native speech, converging on the observation that native disfluencies may aid the listener in comprehension (Arnold et al., 2007; Corley et al., 2007). The non-arbitrary distribution of native disfluencies lead listeners to anticipate reference to a more complex or dispreferred object, following a disfluency (Arnold et al., 2007, 2004; Barr & Seyfeddinipur, 2010). Chapter 4 will extend the understanding of the perceptual effects of disfluencies to the study of L2 speech. The experiments in Chapter 4 will test whether the more irregular patterns of non-native disfluency production lead listeners to attenuate the effect of disfluencies on prediction. For this, we target attribution of

disfluencies to difficulty in formulation (i.e., lexical access) because it is at this particular stage in speech planning that native and non-native speakers diverge. The experiments adopt an adapted version of the methodology of Arnold et al. (2007): participants in an eye-tracking experiment will be presented with pictures of high-frequency (e.g., a hand) and low-frequency objects (e.g., a sewing machine) and with fluent and disfluent spoken instructions (e.g., 'Click on uh.. the [target]'). This allows for an investigation into the first research question of this chapter:

> RQ 3A: Do listeners anticipate low-frequency referents upon encountering a disfluency?

It is expected that, upon encountering a native disfluency, there will be more looks to low-frequency objects than to high-frequency objects. This would suggest that listeners attribute the presence of disfluency to speaker trouble in formulation (i.e., lexical retrieval).

Another experiment will then study non-native disfluencies in order to answer the second research question:

> RQ 3B: Do native and non-native disfluencies elicit anticipation of low-frequency referents to the same extent?

In this second experiment, we will present listeners with L2 speech with a strong foreign accent. If, due to their more irregular distribution, non-native disfluencies are less informative of the word to follow (compared to native disfluencies), we expect to find attenuation of the effect of disfluencies on prediction. In this fashion, it will be investigated whether listeners flexibly adapt their predictive strategies to the (non-native) speaker at hand.

1.5.4 Chapter 5

Finally, Chapter 5 will study the effect of native and non-native disfluencies on attention. It has been argued that the beneficial effects of disfluencies on prediction (Arnold et al., 2007, 2004) and memory (Collard, 2009; Corley et al., 2007; MacGregor et al., 2010) are caused by disfluencies directing the listener's attentional resources (Collard et al., 2008). For instance, in a Change Detection Paradigm, participants were found to be more accurate at detecting substitutions of words that had been encountered in the context of a preceding filled pause. The experiments in Chapter 5 will address the following research question:

> RQ 4: Do native and non-native disfluencies trigger heightened attention to the same extent?

A first experiment aims to replicate the findings from Collard (2009) by testing L1 listeners in a Change Detection Paradigm with a native speaker. A second experiment will subsequently test L1 listeners with L2 speech containing non-native disfluencies. If non-native disfluencies do not trigger listeners' attention as native disfluencies do, this would indicate that listeners are capable of modulating the extent to which disfluencies trigger attention. If, in contrast, both native and non-native disfluencies induce a heightened attention to the following linguistic content, this would suggest that listeners raise their attention in response to disfluency in an automatic fashion without taking the speaker identity into account. Thus, Chapter 5 explores the role of attention in disfluency processing.

1.6 Reading guide

The main chapters (Chapters 2 - 5) of this dissertation have been written as individual papers: each chapter can be read on its own. As a result, there will be some overlap in the method sections and literature overviews. An adapted version of Chapter 2 has been published in the journal *Language Testing*, an adapted version of Chapter 3 has been accepted for publication in the journal *Language Learning*, and an adapted version of Chapter 4 is under review for publication in another journal. The results of various chapters have been presented at international conferences such as *The European Second Language Association* (Stockholm, 2011; Amsterdam 2013), *The European Association for Language Testing and Assessment* (Innsbruck, 2012), *The 11th International Symposium of Psycholinguistics* (Tenerife, 2013), *New Sounds* (Montreal, 2013), and *Architectures and Mechanisms for Language Processing* (Marseille, 2013).

CHAPTER 2

What makes speech sound fluent? The contributions of pauses, speed and repairs[1]

2.1 Introduction

The level of oral fluency of non-native (L2) speakers is an important measure in assessing a person's language proficiency. It is often examined using professional tests (e.g., TOEFL iBT) which may have lasting effects on a person's life in the non-native cultural environment (such as employment or university admission). Therefore, researchers have attempted to unravel the different factors that influence fluency ratings. Two different interpretations of the notion 'fluency' have been distinguished by Lennon (1990): fluency in the broad and in the narrow sense. Fluency in a broad sense is most often used in everyday life when someone claims to be 'fluent' in four languages. In this setting, speaking a language fluently may refer to error-free grammar, a large vocabulary and/or native-like pronunciation. Fluency in the broad sense is equivalent to overall speaking proficiency (Chambers, 1997) and has been further categorized in Fillmore (1979). In contrast, fluency in a narrow sense is a component of speaking proficiency. This sense is often encountered in oral examinations:

[1] An adapted version of this chapter has been published in the journal *Language Testing* as: Bosker, H.R., Pinget, A.-F., Quené, H., Sanders, T.J.M., & De Jong, N.H. (2013) What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing 30* (2), 157-175.

apart from grammar and vocabulary, the flow and smoothness of the speech is also assessed. Fluency in this sense has been defined as an "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and smoothly" (Lennon, 1990, p. 391) and it is this narrow sense that we are concerned with here.

Segalowitz (2010) has, more recently, approached fluency from a cognitive perspective. He argues that sociolinguistic (social context), psycholinguistic (the neurocognitive system of speech production) and psychological (motivation) factors interlinked in a dynamical system all contribute to the level of fluency. Three facets of fluency are distinguished, namely cognitive fluency - "the efficiency of operation of the underlying processes responsible for the production of utterances"; utterance fluency - "the features of utterances that reflect the speaker's cognitive fluency" which can be acoustically measured; and perceived fluency - "the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (Segalowitz, 2010, p. 165). Furthermore, measures of utterance fluency (e.g., number and duration of filled and silent pauses, speech rate, number of repetitions and corrections, etc.) may be clustered into three fluency dimensions: breakdown fluency concerns the extent to which a continuous speech signal is interrupted; speed fluency has been characterized as the rate and density of speech delivery, and repair fluency relates to the number of corrections and repetitions present in speech (Skehan, 2003, 2009; Tavakoli & Skehan, 2005).

The present study investigates the separate contributions of breakdown, speed, and repair fluency to perceived L2 fluency. This issue is approached from two perspectives: from the language testing perspective (Experiment 1) and from a cognitive psychological perspective (Experiments 2-4). Many previous studies have looked at factors influencing raters' judgments (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008); the present study is an attempt to extend this body of research by relating subjective fluency ratings of L2 speech to combinations of acoustic measures, specific to each of the three fluency dimensions. In this fashion we intend to determine the relative contributions of the fluency dimensions to perceived L2 fluency (Experiment 1). Once this will have been established, the question why some fluency dimensions contribute more to fluency perception than others will be addressed. To answer this question, we turn to cognitive psychological factors. More specifically, we hypothesize that listeners' general perceptual sensitivity lies at the foundation of fluency perception. A series of experiments aims to establish the relative sensitivity of listeners to pause phenomena (Experiment 2), to the speed of delivery (Experiment 3) and to repair features in speech (Experiment 4). Results of such investigations license a comparison between listeners' sensitivity to speech characteristics and the factors involved in L2 fluency perception. This comparison is expected to shed light on the question why some fluency dimensions contribute more to

fluency perception than others.

The approach of our experiments involves relating utterance fluency (objective phonetic measurements of L2 speech) to perceived fluency (subjective ratings of the same speech). This approach is often used to gain more insight into the acoustic correlates of oral fluency. For instance, Cucchiarini et al. (2002) had teachers rate speech materials obtained from 30 beginning learners and 30 intermediate learners of Dutch. These perceived fluency ratings were found in subsequent analyses to be best predicted by the number of phonemes per second for beginning learners and by the mean length of run for the intermediate learners. Derwing et al. (2004) used novice raters for obtaining perceived fluency judgments. These raters listened to speech materials of 20 beginner Mandarin-speaking learners of English. Significant correlations were found between the fluency ratings and pausing and standardized pruned syllables per second (the total number of syllables disregarding corrections, repetitions, nonlexical filled pauses, etc.). Rossiter (2009) found the number of pauses per second and pruned speech rate to be strong predictors of perceived fluency. Kormos and Dénes (2004) related acoustic measurements from L2 Hungarian speakers to fluency ratings by native and non-native teachers. They found speech rate, mean length of utterance, phonation time ratio (spoken time / total time x 100%) and the number of stressed words produced per minute to be the best predictors of fluency scores.

A closer look into the methodology and results of these studies reveals much diversity. Conceptual considerations have major effects on the studies' designs and results. To illustrate this point, consider the intercollinearity of acoustic measures of speech. Depending on the specificity of speech annotations, the number of available acoustic predictors of speaking fluency may grow very large. The larger the number of acoustic measures that are related to fluency ratings, the larger the chance of confounding the different measures, which would obscure the interpretability of results. For example, the measures speech rate (number of syllables divided by total time including silences) and mean duration of a silent pause both depend on the duration of silent pauses in the speech signal, and therefore, these two measures are interrelated. If a study would find these two measures to be strongly related to fluency ratings, the relative contribution of each measure to perceived fluency remains unclear, due to the intercollinearity of these measures. In order to understand what raters really listen to when evaluating oral fluency, correlations among acoustic measures should also be taken into account. Unfortunately, correlations between fluency measures are often lacking in the literature, even though the degree of intercollinearity of measures may distinguish orthogonal from confounded measures. Reporting correlations between acoustic measures could identify those measures with low intercollinearity, which would aid the interpretability of results. The present study also emphasizes on the degree of intercollinearity of

our measures. More specifically, the distinction between the three fluency dimensions (breakdown, speed and repair fluency) is central to our selection of acoustic measures. Only those measures that do not confound the fluency dimensions will be employed in our regression analyses.

The first experiment of this study was set up to answer a first research question:

> RQ 1A: What are the independent contributions of the three fluency dimensions of utterance fluency (breakdown, speed, and repair fluency) to perceived fluency?

This issue is approached by relating objective acoustic measurements of speech to subjective fluency ratings of that same speech. A group of untrained raters judged the fluency of L2 Dutch speech excerpts. Derwing et al. (2004) already hypothesized that fluency judgments from untrained native-speaker raters are equivalent to those obtained from expert raters, owing to comparable levels of inter-judge agreement. Rossiter (2009) compared fluency ratings from untrained raters with fluency ratings from expert raters and did not find a statistically significant difference between the two groups. Also, Pinget et al. (in press) have recently demonstrated that untrained raters can keep the concept of fluency well apart from perceived accent. The subjective ratings from the untrained raters from Experiment 1 were modeled by three sets of predictors: a set of pause measures, a speed measure and a set of repair measures. Since the discussed literature (e.g., Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009) mainly found speed and pause measures to be related to fluency ratings, it is expected that both breakdown and speed fluency are primary factors influencing fluency ratings. With respect to repair fluency, the literature seems to suggest that there is no relationship between repair fluency and perceived fluency. For instance, Cucchiarini et al. (2002) did not find any relationship between fluency ratings and number of disfluencies (which covers a.o. repetitions and corrections).

Experiment 1 is expected to shed light on RQ 1A by distinguishing the relative contributions of the three fluency dimensions. Finding an answer to RQ 1A raises a second question of why some fluency dimensions contribute more to fluency perception than others. To this end, the psycholinguistic process of speech perception is investigated. One specific cognitive psychological factor possibly underlying fluency perception is targetted, namely listeners' general perceptual sensitivity. Thus the relationship between the sensitivity of listeners to speech characteristics and fluency perception is studied. It is hypothesized that differences in sensitivity to specific speech phenomena may account for differences in correlations between acoustic measures and fluency ratings. More specifically, if, for instance, pause measures can be found to be strongly related

to perceived fluency ratings, the question can be posed whether this might be due to the fact that listeners are in general more sensitive to pause phenomena. If this scenario can be shown to be true, perception then 'paves the way' for rating: the way we perceive speech influences our subjective impression of that speech. If, in contrast, there is an asymmetry between speech features that contribute to fluency perception and the features in speech listeners are most sensitive to (e.g., pause characteristics are well perceived but contribute only little to fluency perception), then perceptual sensitivity is not the only factor determining fluency perception. Listeners, in this scenario, would first perceive the acoustic characteristics of a speaker's speech but then subsequently also weigh their importance for fluency. These considerations result in the formulation of our second research question:

> RQ 1B: How well can listeners evaluate the pause, speed, and repair characteristics in speech?

To answer RQ 1B, three additional experiments were designed. The crucial distinction between the experiments was the set of instructions given to raters. In Experiment 2 the same L2 speech materials from Experiment 1 were used but a new group of raters received different instructions, namely to rate the use of silent and filled pauses. Relating their pause ratings to objective pause measures is expected to reveal to what extent listeners are sensitive to pauses in speech. Experiment 3 had a similar approach, but now another group of raters was instructed to rate the identical L2 speech materials on the speed of delivery of the speech. And in Experiment 4 yet another group of raters received instructions to rate the L2 speech on the use of repairs (i.e., corrections and hesitations). Findings from these latter three experiments allows us to explore whether the different sensitivities of listeners to acoustic speech characteristics (RQ 1B) may account for the relative contributions of fluency dimensions to perceived fluency (RQ 1A).

2.2 Method

Participants Eighty participants, recruited from the UiL OTS participant pool, were paid for participation in one of four experiments. All were native Dutch speakers without any training in language rating and reported normal hearing (Experiment 1: $N = 20$, $M_{age} = 20.20$, $SD_{age} = 1.88$, 1m/19f; Experiment 2: $N = 20$, $M_{age} = 20.65$, SD_{age}=2.70, 2m/18f; Experiment 3: $N = 20$, $M_{age} = 20.35$, $SD_{age} = 2.76$, 2m/18f; Experiment 4: $N = 20$, $M_{age} = 20.74$, $SD_{age} = 1.79$, 4m/16f).

Stimulus description Speech recordings from native and non-native speakers of Dutch were obtained from the 'What Is Speaking Proficiency'-project (WISP) in Amsterdam (as described in De Jong et al., 2012a). Assessment of these speakers' productive vocabulary knowledge resulted in vocabulary scores which were shown to be strongly related to their overall speaking proficiency (De Jong et al., 2012a). Two non-native speaker groups (15 English and 15 Turkish) were matched for their performance on the vocabulary test (Turkish: $M = 68$, $SD = 18$; English: $M = 64$, $SD = 16$; $t(28) = 0.552$, $p = 0.585$). Moreover, 8 native speakers of Dutch were also selected from the WISP corpus. These were included in order to offer raters reference points to which they could compare the non-native items. The native speakers were selected such that their vocabulary scores were closest to the average of all native speakers (average score of native speakers = 106). All speakers had performed eight different computer-administered speaking tasks. These tasks had been designed to cover the following three dimensions in a $2 \times 2 \times 2$ fashion: complexity (simple, complex), formality (informal, formal) and discourse type (descriptive, argumentative). From these eight tasks, three tasks were here selected. These three tasks covered a range of task characteristics and targeted relatively long stretches of speech. In Table 2.1 descriptions of each of the three tasks are given together with the proficiency level according to the Common European Framework of Reference for Languages (CEFR; Hulstijn, Schoonen, De Jong, Steinel, & Florijn, 2012).

Table 2.1: Descriptions of the selected topics.

	CEFR-level	Characteristics	Description
Topic 1	B1	Simple, formal, descriptive	The participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened.
Topic 2	B1	Simple, formal, argumentative	The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.
Topic 3	B2	Complex, formal, argumentative	The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues which one of three alternative plans for building a car park is to be preferred.

In this fashion, the speech materials consisted of 38 speakers performing 3 tasks (= 114 items). Fragments of approximately 20 seconds were excerpted from approximately the middle of the original recordings. Each fragment started at a phrase boundary (Analysis of Speech Unit; Foster, Tonkyn, & Wigglesworth, 2000) and ended at a pause (> 250 ms). The fragments had a sampling frequency of 44100Hz and were scaled to an intensity of 70dB.

Six objective acoustic measures were calculated for each recording (see Table 2.2; and Appendix A for a link to the raw data) based on human annotations of the speech recordings. Confounding the fluency dimensions was avoided so that each measure was specific to one dimension of fluency. For this reason, all frequency measures were calculated using spoken time (excluding silences) instead of total time (including silences). For instance, previous work suggests that the measure mean length of run correlates with raters' perceptions of fluency (Cucchiarini et al., 2002; Kormos & Dénes, 2004), but because this measure is dependent on the number of pauses in speech it actually combines both speed and breakdown fluency. Therefore, this type of measure was not used in the present study. The dimension of speed fluency was represented by one measure: the mean length of syllables (MLS). A log transformation was performed so that the data would more closely approximate the normal distribution. Breakdown fluency was represented by three measures: the number of silent pauses per second spoken time (NSP), the number of filled pauses per second spoken time (NFP) and the mean length of silent pauses (MLP). A log transformation was performed also on this latter measure for the same reasons as above. These three measures were selected, since we wanted to have separate measures for the number and the duration of silent pauses, and since we wanted to make the distinction between filled and silent pauses. Finally repair fluency was represented by two measures: the number of repetitions (NR) and the number of corrections (NC) per second spoken time. All measures have the same polarity: the higher a value, the less fluent the fragment. The pause exclusion criterion was set at 250 ms, because pauses shorter than 250 ms have been classified as 'micro-pauses' (Riggenbach, 1991) which are irrelevant for calculating measures of fluency (De Jong & Bosker, 2013).

Design and procedure of Experiment 1 The speech fragments of approximately 20 seconds long were presented to participants using the FEP experiment software (version 2.4.19; Veenker, 2006). Participants listened to stimuli over headphones at a comfortable volume in sound-attenuated booths. Written instructions, presented on the screen, instructed participants to judge the speech fragments on overall fluency. In order to avoid the interpretation of fluency in the broad sense (i.e., overall speaking proficiency), participants were instructed not to rate the items in this broad interpretation. In contrast, par-

Table 2.2: List of six selected acoustic measures

Dimension	No.	Acoustic measures	Calculation
Speed	1	Mean length of syllables (MLS)	Log(spoken time / number of syllables)
Breakdown	2	Number of silent pauses (NSP)	Number of silent pauses / spoken time
	3	Number of filled pauses (NFP)	Number of filled pauses / spoken time
	4	Mean length of silent pauses (MLP)	Log(sum of silent pause durations / number of silent pauses)
Repair	5	Number of repetitions (NR)	Number of repetitions / spoken time
	6	Number of corrections (NC)	Number of corrections / spoken time

ticipants were asked to base their judgments on i) the use of silent and filled pauses, ii) the speed of delivery of the speech and iii) the use of hesitations and/or corrections (and not on grammar, for example; see Appendix A). Following the instructions but prior to the actual rating experiment six practice items were presented so that participants could familiarize themselves with the procedure. When participants asked questions to the experimenters, no instructions other than the written instructions were supplied to the participants by the experimenters. There were three different pseudo-randomized ordered lists of the stimuli and three reversed versions of these lists, resulting in six different orders of items. Each session lasted approximately 45 minutes. Participants were allowed to take a brief pause halfway through the experiment. Participants rated the speech fragments using an Equal Appearing Interval Scale (EAIS; Thurstone, 1928). This scale was composed of 9 stars with labeled extremes ("not fluent at all" on the left; "very fluent" on the right; see Appendix A). Above each rating scale a question summarized the rating instructions. At the end of each session the participant filled out a short questionnaire which inquired about attitudes towards and exposure to L2 speech, the factors which the participants themselves thought had influenced them in their rating task (e.g., pauses, speed, repairs, grammar, vocabulary, etc.), and personal details.

Design and procedure of Experiment 2 The speech materials used in the second experiment were identical to those in Experiment 1. A new group of 20 raters participated in this second experiment. The procedure of this experiment was identical to Experiment 1, but crucially the instructions given to these new raters were altered. Participants in Experiment 2 were asked to rate the speech for the use of silent and filled pauses. The instructions to partici-

pants in Experiment 2 were modeled on those used for Experiment 1 (i.e., the introduction, specific formulations and the definitions of pause phenomena; see Appendix A) but no reference was made to the notion of 'fluency'.

Design and procedure of Experiment 3 The speech materials and procedure of the previous experiments were used again for the third experiment. A new group of raters was instructed to rate the L2 speech with the instructions to base their judgments on the speed of delivery of the speech. The literal instructions were modeled on Experiment 1 such that certain terms and the definition of 'speed of delivery' were identical across experiments but without mentioning the term 'fluency' (see Appendix A).

Design and procedure of Experiment 4 In the fourth experiment another group of raters was instructed to rate the same L2 speech materials on the use of hesitations and corrections. Again, definitions of repair phenomena were identical to Experiment 1 but no reference was made to the notion of 'fluency' (see Appendix A).

2.3 Results

Acoustic analysis of stimulus materials First, the non-native speech materials were analyzed (no analysis was performed on (ratings of) native fragments). The intercollinearity of the acoustic measures was investigated through Pearson's r correlations between acoustic measures, in Table 2.3. The correlation measures reported in Table 2.3 allow a comparison between acoustic measures within and across dimensions of fluency. Correlations within fluency dimensions were only possible to analyse for breakdown and repair fluency since speed fluency was represented by one single measure. Within breakdown fluency only one statistically significant correlation was found, namely a weak correlation between NSP and NFP ($r = -0.248$). Within repair fluency, the correlation between the two measures was not statistically significant. Correlations across fluency dimensions primarily concerned weak to moderate correlations with the speed fluency measure MLS, and a correlation between NSP and NC was also found. The relationship between acoustic measures within fluency dimensions was similar to the relationship between acoustic measures across fluency dimensions.

In addition, correlations between single acoustic measures and the fluency ratings were calculated (see Table 2.3). The highest observed correlation was between the speed measure mean length of syllables and the fluency ratings ($r = -0.742$). In order to investigate the contribution of fluency dimensions to perceived fluency, additional analyses were performed.

Table 2.3: Correlations (Pearson's r) between acoustic measures and between acoustic measures and fluency ratings.

Acoustic measure	Speed	Breakdown			Repair		Fluency ratings
	MLS	NSP	NFP	MLP	NR	NC	
Mean length of syllables (MLS)	1						-0.742***
Number of silent pauses (NSP)	0.330**	1					-0.422***
Number of filled pauses (NFP)	0.308**	-0.248*	1				-0.154
Mean length of silent pauses (MLP)	0.152	-0.096	-0.168	1			-0.470***
Number of repetitions (NR)	0.292**	0.037	0.188	0.034	1		-0.348***
Number of corrections (NC)	0.102	0.216*	-0.037	-0.088	0.012	1	-0.241*

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Results Experiment 1 Each item in Experiment 1 was rated by 20 judges. The extent to which raters in Experiment 1 agreed with each other was high (Cronbach's alpha coefficient: 0.97). In order to relate these subjective ratings of each item to the objective acoustic properties of that item, a method of collapsing these 20 ratings for each item was required. Many previous fluency studies take the mean of the collected ratings for each item, thereby disregarding such confounding factors as individual differences between raters, for instance, or effects of presentation order. Our analyses were performed in two consecutive steps. The first step involved correcting the fluency ratings for these confounding factors using Best Linear Unbiased Predictors (Baayen, 2008, p. 247), which resulted in corrected estimates of the raw fluency ratings. The correction procedure was performed using Linear Mixed Models (cf. Baayen, Davidson, & Bates, 2008; Quené & Van den Bergh, 2004, 2008) as implemented in the `lme4` library (Bates, Maechler, & Bolker, 2012) in R (R Development Core Team, 2012). Thus we controlled for three confounding factors: Order (fixed effect) testing for general learning or fatigue effects; Rater (random effect) testing for individual differences between raters; and OrderWithinRaters (random effect) testing for individual differences in order effects. Simple models, containing one or two of these predictors, were compared to more complex models that contained one additional predictor. In order to allow such comparisons of models in our analysis, coefficients of models were estimated using the full maximum

likelihood criterion (Hox, 2010; Pinheiro & Bates, 2000). Likelihood ratio tests (Pinheiro & Bates, 2000) showed that the most complex model proved to fit the data of Experiment 1 better than any simpler model. This optimal model showed significant effects of Rater, of Order (raters became harsher to the L2 speech as the experiment progressed) and of OrderWithinRaters (the order effect differed among individual raters). This optimal model was used to predict estimates of the fluency ratings. This was the first step of the investigative procedure reported here. All subsequent analyses were performed on these corrected estimates instead of on averages (see Appendix A for a link to the raw data).

The second step involved relating objective acoustic measures to these corrected estimates of the fluency ratings. Multiple linear regression analyses were performed in order to explore to what extent a set of objective acoustic measures could explain the variance of the (estimated) fluency ratings, gauged by the adjusted R^2.

Because the present study is primarily concerned with the contributions of fluency dimensions, and not of single acoustic measures, predictors in the multiple linear regression models were sets of acoustic measures and not single acoustic measures. All measures were centralized to their median value. In Table 2.4 six different models of the fluency judgments are summarized. Because effects of the L1 language (English vs. Turkish) and of the different speaking tasks were not statistically significant, these factors will be ignored in the present multiple linear regression analyses.

Table 2.4: Models predicting the fluency estimates of Experiment 1 using acoustic measures.

Model	Predictors	Adjusted R^2	Significance testing
(1)	NSP*NFP*MLP (breakdown)	0.5917	
(2)	MLS (speed)	0.5449	
(3)	NR+NC (repair)	0.1583	
(4)	NSP*NFP*MLP (breakdown) + MLS (speed)	0.7825	Model 4 vs. 1: $F(1, 82) = 73.793$, $p < 0.001$
(5)	NSP*NFP*MLP (breakdown) + NR+NC (repair)	0.6804	Model 5 vs. 1: $F(1, 81) = 12.523$, $p < 0.001$
(6)	NSP*NFP*MLP (breakdown) + MLS (speed) + NR+NC (repair)	0.8378	Model 6 vs. 4: $F(1, 80) = 15.004$, $p < 0.001$

Firstly, three models (1-3) were built with predictors from only one of the fluency dimensions. Model (1) included the three acoustic measures specific to

breakdown fluency: NFP, NSP and MLP. A comparison between a model with no interactions and a model with three two-way interactions demonstrated that the model with the three two-way interactions had a significantly stronger explanatory power and therefore these three two-way interactions were included in all subsequent models. This model resulted in an adjusted R^2 of 0.5917. Model (2) predicted fluency ratings using the speed measure MLS as predictor, and it resulted in an adjusted R^2 of 0.5449. Model (3) had the repair fluency measures, NC and NR, as predictors of perceived fluency (adjusted $R^2 = 0.1583$).

Seeing that model (1) with breakdown fluency measures as predictors explained the largest part of the variance of the fluency ratings, we tested whether additional contributions of speed fluency and of repair fluency added to the predictive power of the model. Model (4) additionally contained the acoustic measure specific to speed fluency, MLS (adjusted $R^2 = 0.7825$), and model (5) also included the repair fluency measures, NC and NR (adjusted $R^2 = 0.6804$). As evidenced by the higher adjusted R^2 values relative to model (1) and by the statistical comparisons of models, both models improved the explanatory power of model (1) with model (4) yielding a higher adjusted R2 than model (5). Finally, the most complex model (6) which included all fluency dimensions as predictors yielded the highest adjusted R^2 of 0.8378.

When comparing these results with the responses from the participants to the questions in the post-experimental questionnaire, it was found that participants themselves reported to have been mainly influenced by pauses ($n = 19$) and speed ($n = 15$) and less so by repetitions and corrections ($n = 12$).

Results Experiments 2-4 In Experiments 2-4 all stimulus material was kept constant, but new groups of raters received different instructions, namely to rate the speech on the use of silent and filled pauses (Experiment 2), on the speed of delivery (Experiment 3) and on the use of repetitions and corrections (Experiment 4). Raters within the separate experiments strongly agreed as evidenced by high Cronbach's alpha coefficients calculated using the raw ratings: 0.95 (Experiment 2); 0.96 (Experiment 3); 0.94 (Experiment 4). The analyses of the different experiments again involved two steps. Firstly, the raw ratings were corrected for confounding random effects. It was established that for all experiments the most complex Linear Mixed Model, which included Order, Rater and OrderWithinRaters as predictors, proved to fit the raters' data the best. The estimates resulting from these models were taken as dependent variable in the second step of the analyses (see Appendix A for a link to the raw data). This second step involved modeling the subjective estimates of each experiment by objective measures from the appropriate fluency dimension (i.e., speed ratings by speed measures, pause ratings by pause measures, and repair

ratings by repair measures). As given in Table 2.5, model (7), predicting subjective pause ratings using pause measures, was observed to have the highest adjusted R^2 value (0.6986) of the three analyses. Model (8) and (9) perform worse than model (7) and explain almost the same amount of variance. The responses from the participants to the questions in the post-experimental questionnaire did not reveal any particular pattern, except that each group said to have been mainly influenced by the 'relevant' acoustic factor (e.g., pause raters by pauses, speed raters by speed, repair raters by repairs).

Table 2.5: Models predicting the estimates of Experiments 2-4 using acoustic measures.

Model	Dependent variable	Predictors	Adjusted R^2
(7)	Pause ratings from Experiment 2	NSP*NFP*MLP	0.6986
(8)	Speed ratings from Experiment 3	MLS	0.5287
(9)	Repair ratings from Experiment 4	NR+NC	0.5452

Subjective ratings as predictors for fluency ratings The data resulting from Experiments 2-4 allow for an additional analysis of the results of Experiment 1. Using the same materials, the subjective fluency ratings from Experiment 1 were predicted by the subjective ratings of specific speech characteristics from Experiments 2-4, see Table 2.6. These results show that most of the variance of the fluency judgments may be predicted by subjective pause ratings. The model with the 'best fit' was the most complex model (15), with the ratings of all three subjective dimensions included as predictors.

Table 2.6: Models predicting the fluency estimates of Experiment 1 using subjective ratings.

Model	Predictors	Adjusted R^2	Significance testing
(10)	Pause estimates	0.8523	
(11)	Speed estimates	0.7829	
(12)	Repair estimates	0.2735	
(13)	Pause estimates + Speed estimates	0.8923	Model 13 vs. 10: $F(1, 87) = 34.626$, $p < 0.001$
(14)	Pause estimates + Repair estimates	0.8807	Model 14 vs. 10: $F(1, 87) = 21.873$, $p < 0.001$
(15)	Pause estimates + Speed estimates + Repair estimates	0.9208	Model 15 vs. 13: $F(1, 86) = 31.400$, $p < 0.001$

2.4 Discussion

This study investigated the contributions of three dimensions of fluency (breakdown, speed and repair fluency) to perceived fluency ratings. In Experiment 1, untrained raters evaluated L2 speech items with regards to fluency, with the aim of establishing the contributions of the different fluency dimensions to fluency perception (RQ 1A). Sets of acoustic measures relating one of three fluency dimensions were included in models predicting the subjective fluency ratings. Cross-correlations between the speech measures demonstrated that both within and across fluency dimensions our speech measures were largely independent. This low intercollinearity aided the interpretation of other analyses. De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012b) also report on correlations between acoustic measures within and across fluency dimensions. A comparison reveals that the relationship between measures that theoretically cluster together within fluency dimensions show, in both studies, no stronger correlations amongst each other than measures across fluency dimensions do. Together with De Jong et al. (2012b) we argue that measures from the same fluency dimension might be caused by the same cognitive problems in the speech production process. Where one speaker would use a silent pause to win time, another might resort to the use of filled pauses, resulting in low correlations between the two measures. Future research into the specific function of disfluencies in (L1 and L2) natural speech will have to address this issue.

Having established that the acoustic measures used in our analyses did not confound the fluency dimensions, we turn to RQ 1A. Comparisons between fluency models revealed that all three dimensions play a role in fluency perception and none of these dimensions should be disregarded. Still, breakdown fluency explained the largest part of the variance in subjective fluency ratings, closely followed by speed fluency. Strong correlations between pause and speed measures and fluency ratings as reported in previous literature (Derwing et al., 2004; Rossiter, 2009) support this major role of breakdown and speed fluency. In addition, correlations between single acoustic measures and the fluency ratings suggest that the major role of breakdown fluency is primarily due to the effect of (the duration and the number of) silent pauses rather than filled pauses.

The second research question sought to find a possible explanation for this finding by investigating the perceptual sensitivity of listeners. It was argued that differences in perceptual sensitivity of listeners to certain speech characteristics might account for different contributions of fluency dimensions to fluency perception. The results from Experiments 2-4 would then mirror those from Experiment 1: breakdown and speed fluency should be well perceived but repair fluency should be perceived less accurately. RQ 1B studied the sensi-

tivity of listeners to the three fluency dimensions in three experiments that collected ratings of pausing, speed and repairs. As expected, the ratings from Experiment 2 on pausing were, of all three fluency dimensions, best predicted by acoustic measures as evidenced by the highest adjusted R^2 value (Table 2.5). Since the subjective pause ratings were well accounted for by the objective acoustic properties of the speech, we argue that listeners are apparently most sensitive to pause characteristics of speech. Listeners are also sensitive to speed characteristics of speech, though less sensitive as compared to pause features. Surprisingly, listeners were also found to be sensitive to speech repairs. In fact, they are approximately as sensitive to speed features as they are to repairs. If perceptual sensitivity of listeners were the only factor determining the relative contributions of fluency dimensions to fluency perception, then we would, based on the results from Experiment 2-4, expect to have found a larger contribution of repair measures to the perception of fluency in Experiment 1. Apparently, listeners weigh the perceived speech characteristics on their importance for fluency judgments.

The first research question was approached in Experiment 1 by relating objective acoustic measurements from three dimensions of fluency to subjective ratings. Additional support for the findings from Experiment 1 was found by relating the subjective perception of the three fluency dimensions (Experiment 2-4) to subjective ratings of fluency (Experiment 1). These supplementary models substantiated the findings from previous models: all three dimensions are involved in fluency perception but breakdown and speed fluency are most strongly related to fluency perception.

Based on the results from Experiment 1 it is evident that repair phenomena, though they are well perceived, contribute only little to fluency perception. A possible account for this might be that our repair measures were not sensitive enough to expose the contribution of repair fluency to fluency perception. For instance, it has been proposed to distinguish between error repairs - repairing errors of linguistic form; and appropriateness repairs - presenting a new or rephrased message (Kormos, 1999; Levelt, 1983). Our current repair measures may have lacked the precision to adequately study the contribution of repair fluency. In addition, our repair measures only captured the frequency of occurrence of corrections and repetitions. As such, these measures are insensitive to the extent of repairs (e.g., the number of extraneous words involved). Several quick repetitions of single words may be perceived as less obstructive than lengthy garbles requiring major backtracking. However, despite the shortcomings of our repair measures, there is to our knowledge no evidence in the literature for a relation between speech repairs and fluency perception. Cucchiarini et al. (2002) could not find any relationship between repairs and fluency perception. Repetitions also seem to differ from other types of disfluencies with respect to the online processing of speech. MacGregor, Corley, and Donaldson

(2009) did not find an N400 attenuation effect for repetitions or any memory effect, where these effects were established for filled pauses (Corley et al., 2007). Gilabert (2007) takes corrections in speech primarily as a measure of accuracy rather than fluency since corrections both denote attention to form and an attempt at being accurate. Apparently, there is no consensus on the function repairs have in speech perception. The contribution of repair phenomena to fluency perception clearly deserves more attention.

One of the limitations of the current study concerns the character of the analyses. Relationships between sets of acoustic measures and fluency perception were gauged by means of correlational analyses. One must be careful not to automatically interpret the relationships found as causal relationships (i.e., "the fluency rating of item A was higher than item B because of the larger number of pauses in item B"). The present study cannot decide on the nature (e.g., direct or indirect) of the relationships that were found. Causal relationships can only be laid bare when one specific factor of interest is manipulated and all other interacting factors are kept constant (ceteris paribus). Future research, involving manipulating speech characteristics in different dimensions and studying its effect on fluency perception, will have to illuminate the nature of the relationships found in the present study. Interesting in this respect would be to study effects both in L2 fluency and in L1 fluency. The current study only studied L2 fluency and therefore it remains to be shown whether pause and speed characteristics of speech also play a large role in L1 fluency perception. Based on the fact that we have shown that listeners are perceptually very sensitive to pause and speed features of speech, it may be hypothesized that a similar hierarchy of fluency dimensions may be found for L1 fluency.

The fact that we have demonstrated breakdown and speed fluency to be most strongly related to fluency perception has implications for language testing practice. With respect to automatic fluency assessment, for instance, our results indicate that speed and breakdown measures resemble human fluency perception to a very large extent. This observation corroborates the use of such measures in automatic fluency assessment. Also, from the perspective of the language learner, apparently those L2 speakers that manage to speak relatively fast with only minor pauses are more leniently judged by fluency raters than speakers who never repair at the cost of the speed of delivery and pausing. This observation may lead L2 speakers to prioritize improvements to the flow of their speech, rather than the absence of overt repairs.

2.5 Conclusion

The present study investigated the contribution of three dimensions of fluency (breakdown, speed and repair fluency) to the perception of fluency. Based on

comparisons between models of subjective fluency ratings, we conclude that the dimensions of breakdown and speed fluency are most strongly related to fluency perception. From an investigation into the perceptual sensitivity of listeners to different speech characteristics, it was established that perceptual sensitivity is not the only factor deciding on which dimensions contribute to fluency perception. Apparently, listeners weigh the importance of the perceived dimensions of fluency to come to an overall judgment. This importance of fluency dimensions is, then, not only determined by which speech characteristics are well perceived by the listener.

CHAPTER 3

Perceiving the fluency of native and non-native speech[1]

3.1 Introduction

This chapter is concerned with the difference in the perception of fluency in native and non-native speech. Fluency has been termed "an automatic procedural skill" (Schmidt, 1992) that encompasses the notion of "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language" (Lennon, 2000, p. 20). Lennon (1990) has distinguished between fluency in the broad sense, that is, global speaking proficiency, and fluency in the narrow sense, that is, the "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (Lennon, 1990, p. 391). Segalowitz (2010) distinguishes between three facets of fluency, namely *cognitive fluency* – "the efficiency of operation of the underlying processes responsible for the production of utterances"; *utterance fluency* – "the features of utterances that reflect the speaker's cognitive fluency" which can be acoustically measured; and *perceived fluency* – "the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (Segalowitz, 2010, p. 165). In this study, we are concerned with the relationship between utterance fluency

[1] An adapted version of this chapter has been accepted for publication in the journal *Language Learning*: Bosker, H.R., Quené, H., Sanders, T.J.M., & De Jong, N.H. (in press) The perception of fluency in native and non-native speech. *Language Learning*.

and perceived fluency. Despite the fact that the aforementioned definitions of fluency may apply to both native and non-native speech, fluency assessment has thus far mostly (if not exclusively) aimed at non-native speakers. Native speakers are supposedly perceived as fluent by default even though they, too, produce disfluencies such as *uhm*'s, silent pauses and repetitions. In fact, it is estimated that 6 in every 100 words is affected by disfluency (Fox Tree, 1995) and various factors have been found to influence native disfluency production, including speaker gender, speaker age, conversational topic, planning difficulty, etc. (Bortfeld et al., 2001). Therefore, the current chapter compares the way native and non-native fluency characteristics are weighed by listeners.

The production of non-native disfluencies has been widely studied. Producing fluent speech is an important component of speaking proficiency for non-native speakers as defined in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The descriptors in the global scale (p. 24) state that speakers at level B2 can communicate "with a degree of fluency"; at level C1, speakers can express themselves "fluently", and at level C2, "very fluently". In language testing practice, human raters frequently assess non-native speakers' fluency levels (e.g., Iwashita et al., 2008). Many studies have investigated the acoustic fluency characteristics of non-native speakers. The literature ranges from child L2 learners (Trofimovich & Baker, 2007) to very advanced L2 speakers (Riazantseva, 2001). Non-native speech is reported to contain more disfluencies than native speech (e.g., Cucchiarini et al., 2000) and non-native speakers become more fluent as their proficiency in the non-native language advances (e.g., Freed, 2000; Towell et al., 1996). De Jong, Groenhout, Schoonen, and Hulstijn (2013) have argued that the fluency characteristics of one's L2 speech are strongly related to those in the talker's L1 (cf. Segalowitz, 2010). Both a persons individual traits and the speakers non-native proficiency level define the speakers L2 *cognitive fluency*, with consequences for the fluency characteristics of the speech signal (*utterance fluency*). The utterance fluency of a speaker (i.e, the number of silent pauses per minute, the number of filled pauses, repetitions, corrections, etc.) affects, in turn, the fluency impression that listeners have of a particular speaker (*perceived fluency*).

There have been numerous studies investigating the subjective fluency level of non-native speakers (e.g., Cucchiarini et al., 2000, 2002; Derwing et al., 2004; Freed, 2000; Ginther, Dimova, & Yang, 2010; Kormos & Dénes, 2004; Mora, 2006; Rossiter, 2009; Wennerstrom, 2000). All these studies involve relating measures of perceived fluency (listener ratings, typically involving 7- or 9-point scales) to utterance fluency (temporal speech measures) in order to assess the relative contributions of different speech characteristics to fluency perception. These studies indicate that temporal measures alone can account for a large amount of variance in perceived fluency ratings. Rossiter (2009) reports a correlation of $r = 0.839$ between subjective fluency ratings and pruned num-

ber of syllables per second (the total number of syllables minus disfluencies). She also compared ratings from untrained and expert fluency raters and did not find a statistically significant difference between the two groups. Derwing et al. (2004) used novice raters to obtain perceived fluency judgments. These raters listened to speech materials of 20 beginner Mandarin-speaking learners of English. Derwing et al. (2004) found that pausing and pruned syllables per second together accounted for 69% of the variance of their fluency ratings. Kormos and Dénes (2004) related acoustic measurements from non-native Hungarian speakers to fluency ratings by native and non-native teachers. They reported a correlation of $r = 0.87$ between the measure of speech rate and subjective fluency ratings. Cucchiarini et al. (2002) had teachers rate spontaneous speech materials obtained from non-native speakers of Dutch. They found a correlation of $r = 0.65$ between the mean length of runs (mean number of phonemes between silent pauses) and the perceived fluency of spontaneous speech.

These studies suggest that temporal factors are major contributors to fluency judgments. However, many researchers have raised the question whether non-temporal factors, such as grammatical accuracy, vocabulary use, or foreign accent, should also be considered as influencing fluency judgments (Freed, 1995; Lennon, 1990). Rossiter (2009) notes that subjective ratings of fluency, in her study, were influenced by non-temporal factors as well (on the basis of qualitative analyses of rater comments). The most important factor in this respect was learners' pronunciation of the non-native language. More recently, a quantitative study by Pinget et al. (in press) has tackled the relationship between perceived fluency and perceived accent. This study suggests that raters can keep the concept of fluency well apart from perceived foreign accent. Fluency ratings and accent ratings of the same speech samples were found to correlate only weakly ($r = -0.25$) and, moreover, acoustic measures of accent did not add any explanatory power to a statistical model of perceived fluency. This suggests that, although the contribution of non-temporal factors to perceived fluency should not be ignored, these non-temporal factors likely play only a minor role.

Taking all the evidence together, studies targeting non-native fluency perception converge on the view that acoustic measures of fluency can account for fluency ratings to a large extent. However, as noted, the emphasis of the aforementioned studies is on the level of fluency of *non-native speakers*. Studies exploring the relationship between utterance fluency and perceived fluency of native speakers are rare. Native speakers are supposedly perceived as fluent by default (Davies, 2003; Riggenbach, 1991). Nevertheless, individual differences between native speakers in the production of disfluencies have been reported (Bortfeld et al., 2001). The psychological literature has primarily studied disfluency as a window into different stages of speech planning (e.g., Goldman-Eisler, 1958a, 1958b; Levelt, 1989; Maclay & Osgood, 1959). The study of

speech pathology and speech therapy has primarily focused on the factors that influence (atypical) disfluency production (Christenfeld, 1996; Panico, Healey, Brouwer, & Susca, 2005; Susca & Healey, 2001). However, it is unclear how these disfluencies in native speech are perceived by the listener. From the field of social psychology we know that listeners constantly make inferences about speakers based on the (non-linguistic) content of speech, engaging in what is called person or speaker perception (Krauss & Pardo, 2006). Listener attributions may range from social status (Brown, Strong, & Rencher, 1975) and emotion (Scherer, 2003) to metacognitive states (Brennan & Williams, 1995) and even to physical properties of a speaker (Krauss, Freyberg, & Morsella, 2002). Nevertheless, it is as yet unknown how the fluency characteristics of native speech contribute to the perception of a native speaker's fluency level. The few studies that have included native speech in their fluency research report that natives are consistently rated higher than non-natives (Cucchiarini et al., 2000) and that they also produce fewer disfluencies than non-natives do (Cucchiarini et al., 2000). Ginther et al. (2010) report higher overall oral proficiency for native speakers as measured by an oral English proficiency test as compared to non-native speakers. From these studies, we cannot gather how listeners weigh native and non-native fluency characteristics. In order to gain more insight into the perception of fluency in native and non-native speech, the current work addresses the following research question:

> RQ 2: Do listeners evaluate fluency characteristics in the same way in native and non-native speech?

One could propose to address this question through correlational analyses (cf. Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009), which would involve collecting subjective fluency judgments of native and non-native speech, collecting objective acoustic measurements from native and non-native speech, and then statistically testing to what extent the acoustic measures can account for the subjective ratings. This correlational approach is, however, unsuitable for the comparison of the perception of L1 and L2 speech, because native and non-native speech differs in many respects. The hypothetical observation that silent pauses play a large role when rating non-native fluency, compared to rating native fluency, could simply be accounted for by a difference in pause incidence in native and non-native speech (rather than by a difference in relative weight of pausing). Therefore, a comparison between native and non-native fluency perception is only viable when native and non-native speech samples have been matched for the acoustic dimensions under study.

In order to circumvent this problem, we propose a different method for investigating the contribution of acoustic variables to fluency judgments. We propose to use experiments with acoustic manipulations of the speech signal so

as to ascertain that observed effects in fluency judgments may be directly attributed to particular fluency characteristics (cf. Munro & Derwing, 1998, 2001, who used phonetic manipulations to study perceived accent). The advantage of this method is that it becomes possible to compare native and non-native fluency perception. For instance, we may compare how the same modification of silent pauses in native and non-native speech affects the perception of fluency. If different fluency ratings are given to two speech samples differing in a single manipulated phonetic property, then this perceptual difference may be reliably attributed to the minimal acoustic difference between the samples. This experimental method has the additional advantage that separate contributions of multiple acoustic factors can be investigated. Thus, the effect of one acoustic property on fluency judgments can be singled out through the use of phonetic manipulations targeting the disfluencies in the speech whilst keeping all other possibly interacting factors constant. Even different properties of one and the same acoustic phenomenon can thus be studied, such as the number and the duration of silent pauses. It is difficult to disentangle the contributions of these properties of silent pauses to fluency ratings using correlational analyses. The current approach could thus shed light on differential effects of two pause properties by manipulating pause duration while keeping the number of pauses constant.

The present study reports on two experiments that aim to answer the research question above by studying two different fluency dimensions, namely pausing and speed characteristics of native and non-native speech. Both experiments make use of phonetic manipulations in native and non-native speech. In Experiment 1, the silent pauses present in native and non-native speech were manipulated. In Experiment 2, the speed of native and non-native speech was modified. In our analyses, the main objective was to determine whether or not our manipulations affect fluency ratings of native and non-native speech in a similar fashion.

Two possible hypotheses can be proposed with respect to the distinction between native and non-native fluency. The effects of phonetic manipulations could be similar across native and non-native fluency perception, such that both are equally affected by phonetic manipulations. The literature on non-native fluency perception has shown that fluency judgments depend on the disfluencies in the speech signal (e.g., Cucchiarini et al., 2000, 2002; Derwing et al., 2004; Rossiter, 2009). But native speech also contains disfluencies and manipulating these might have similar effects on fluency ratings as compared to non-native disfluencies.

Alternatively, manipulating characteristics of fluency in the speech signal may also have differential effects on the perception of native and non-native fluency. For example, since natives are proficient in their native language, they are generally perceived as fluent. Therefore, the addition of disfluency char-

acteristics may affect native speech to a lesser extent than non-native speech. The same line of reasoning could also support the opposite prediction: since natives are generally perceived as fluent, the added disfluencies may - in the perception of listeners - stand out more than non-native disfluencies. Therefore, our manipulations could also affect native speech to a larger extent than non-native speech.

The production literature (e.g., Davies, 2003; Skehan, 2009; Skehan & Foster, 2007; Tavakoli, 2011) seems to suggest that native and non-native fluency characteristics may be weighed differentially by listeners. For instance, Skehan and Foster (2007) observed that native speakers have a different pause distribution compared to non-native speakers. Differences in the position of pauses may lead to differential perception of pauses in native and non-native speech. It has even been argued that disfluencies in native speech can help the listener. For instance, eye-tracking data indicate that hesitations may aid the listener in reference resolution. In the study of Arnold et al. (2007), listeners were presented with both a known and a novel visual object on a computer screen. They found that hesitations in the speech signal created an expectation for a novel target word as judged by increased fixations on the novel object. Although research on the role of disfluencies produced by non-natives in listener comprehension of speech is, as yet, still lacking, native disfluencies may differ from non-native disfluencies in their function in speech processing. Non-native disfluencies, for instance, may arise from incomplete knowledge (grammar and/or vocabulary), or insufficient skills (automaticity) in the non-native language and thus hinder native speech processing. This difference in the psycholinguistic source of disfluencies may lead to differences in how listeners judge native and non-native fluency.

3.2 Experiment 1

In Experiment 1, both the duration and the number of silent pauses were independently manipulated. These phonetic manipulations were performed both in native and non-native speech. Native and non-native speech materials were matched for the manipulated dimension. In Experiment 1, this was achieved by matching the native and non-native speech materials for the number of silent pauses. The phonetic manipulations of Experiment 1 involved three pause conditions: speech materials in which silent pauses had been removed, speech materials in which the duration of silent pauses had been altered to be relatively short and speech materials in which their duration was relatively long. We expected that native speech would be rated as being more fluent than non-native speech due to differences between native and non-native speech in fluency characteristics irrespective of the phonetic manipulations (e.g., filled pauses). We

also predicted that fluency would be rated lower when there are more pauses (increasing the number) and/or longer pauses (increasing the duration). We did not make a clear prediction for a possible interaction between the manipulation effects and nativeness. On the one hand, it was possible that the phonetic manipulations would affect ratings of native and non-native fluency in a similar fashion. On the other hand, it was also possible that the phonetic manipulations would have different effects on native fluency perception, as compared to non-native fluency perception (cf. the two introduced hypotheses above).

3.2.1 Method of Experiment 1

Participants Participants were 73 paid members of the UiL OTS participant pool. All were native Dutch speakers who reported to have normal hearing (age: $M_{age} = 20.56$, $SD_{age} = 3.00$; 15m/58f) and who participated with implicit informed consent in accordance with local and national guidelines. A post-experimental questionnaire inquired (amongst other issues) whether they had noticed anything particular about the experiment. In particular, they were asked whether they thought the speech had been digitally edited, and if so, how. In total, 27 participants responded that they thought the stimuli had been edited in some particular way. Individual responses ranged from comments about non-native accents to different amounts of background noise or the censoring of personal details. All responses from participants which could reasonably be interpreted as relevant to pause manipulations were taken as evidence of awareness of the experimental manipulation ($n = 14$; 19%). Data from these participants were excluded from any further analyses. The post-experimental questionnaire also assessed participants' prior experience in teaching L2 Dutch or rating fluency. One participant indicated to have taught L2 Dutch previously and was excluded for this reason. The mean age of the remaining 58 participants was 20.39 years ($SD = 3.15$; 11m/47f).

Stimulus description Speech recordings from native speakers and non-native speakers of Dutch were obtained from the "What Is Speaking Proficiency"-corpus (WISP) in Amsterdam (as described in De Jong et al., 2012a). This corpus was selected because it contains recordings from a large range of native and non-native speakers of Dutch. All speech in the WISP-corpus was collected with signed informed consent from the speakers in accordance with local and national guidelines. All speakers in this corpus had performed computer-administered monologic speaking tasks on eight different topics. These topics had been designed to cover the following three dimensions in a $2 \times 2 \times 2$ fashion: *complexity* (simple, complex), *formality* (informal, formal) and *discourse type* (descriptive, argumentative). For each task, instruction screens provided a picture of the communicative situation and one or several visual-verbal cues

concerning the topic. Participants were informed about the audience they were expected to address in each task and were requested to 'role play' as if they were actually speaking to these audiences. From the eight topics, three topics were selected that covered a range of characteristics and that elicited sufficiently long stretches of speech (approximately 2 minutes). In Table 3.1, descriptions are given of the different topics, together with the proficiency level according to CEFR (Hulstijn et al., 2012).

Table 3.1: Descriptions of the selected topics.

	CEFR-level	Characteristics	Description
Topic 1	B1	Simple, formal, descriptive	The participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened.
Topic 2	B1	Simple, formal, argumentative	The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.
Topic 3	B2	Complex, formal, argumentative	The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues which one of three alternative plans for building a car park is to be preferred.

In total, 10 native speakers and 10 non-native speakers of Dutch were selected. In order to avoid homogeneity in L1 background, non-native speakers from two L1 backgrounds were selected (5 English and 5 Turkish). Proficiency in Dutch was assessed by means of a productive vocabulary knowledge test with 116 items, shown to be strongly related to the speakers' overall speaking proficiency (De Jong et al., 2012a): $M_{L1} = 106$, $SD_{L1} = 5$; $M_{L2} = 69$, $SD_{L2} = 22$ (max=116). Comparing these scores to Hulstijn et al. (2012), we find that our non-native speakers scored approximately at B2 level indicating an intermediate proficiency in Dutch. Their mean length of residence was 7.33 years ($SD = 5.42$) and their mean age of acquisition was 24.9 years ($SD = 3.38$). Fragments of approximately 20 seconds were excerpted from roughly the middle of the original recordings. Thus, 60 speech fragments from 20 speakers talking about three topics were created. All fragments started at a phrase boundary, according to the Analysis of Speech Unit (AS-unit; Foster et al., 2000). Most

of the fragments also ended at a phrase boundary (native: $n = 23$ out of 30; non-native $n = 22$ out of 30), but all fragments ended at a pause (>250 ms).

We attempted to manipulate our native and non-native speech materials in a similar fashion. Therefore, the native and non-native speakers were matched for the number of silent pauses per 100 syllables ($M_{L1} = 6.1$, $SD_{L1} = 2.0$; $M_{L2} = 6.5$, $SD_{L2} = 2.2$; see Appendix B for a link to the raw data).

The excerpted speech fragments served as the basis of our stimulus materials. Each speech fragment was manipulated resulting in three different experimental conditions using Praat (Boersma & Weenink, 2012). The three conditions differed in the manipulations targeting pauses with a duration of more than 250 ms. De Jong and Bosker (2013) have demonstrated that a silent pause threshold of 250 ms leads to acoustic measures that have the highest correlation with L2 proficiency (but see Hieke, Kowal, & O'Connell, 1983).

In the NoPauses condition, all pauses of >250 ms were 'removed' by changing the duration to <150 ms. This was achieved by excising silence in between two extremes at positive-going zero-crossings in the speech signal. The other two conditions were designed on the basis of the 'NoPauses' condition. In the ShortPauses condition, pauses that originally had a duration of >250 ms, were now altered to have a duration of 250-500 ms. This was achieved by adding silence to the NoPauses condition (extracted silent intervals of that particular recording). In the LongPauses condition, pauses of >250 ms were altered to have a duration of 750-1000 ms. We decided on these two duration intervals because research shows that silent pauses of 250-1000 ms are very common in native speech (Campione & Véronis, 2002) and in non-native speech (De Jong & Bosker, 2013). Also, in this fashion, the ShortPauses condition would be clearly distinct from the LongPauses condition with no overlap between the ShortPauses interval of 250-500 ms and the LongPauses interval of 750-1000 ms. Pauses close to the silent pause threshold (i.e., between 150 and 250 ms) were decreased in duration to <150 ms in each of the three conditions. If a speech fragment contained fewer than three pauses of >250 ms, then some pauses of <250 ms were also manipulated such that the number of manipulated pauses per item would add up to at least three. Table 3.2 provides examples of each of the three pause conditions. Note that our phonetic manipulations involved adjustment of silent pauses already present in the original recordings, such that no supplementary silent pauses were added to the speech.

In natural speech, the ratio between inspiration time and expiration time is about 10% inspiration time and 90% expiration time (Borden, Raphael, & Harris, 1994, p. 64-65). Therefore, the silent pauses in the NoPauses condition could not all be excised without impairing the naturalness of our materials. For that reason one pause containing a breath located roughly in the middle of a speech fragment was exempted from manipulations in all conditions (not included in the data shown in Table 3.3).

Table 3.2: Examples of speech fragments on topic 1 from a native and non-native speaker. Silent pause durations (ms) of the three conditions are given as [NoPauses; ShortPauses; LongPauses]. Translations from Dutch to English are provided below each example.

Native speech fragment
uh ik zag een [40; 364; 804] vrouw op de fiets bij een uh stoplicht [54; 352; 910] door een groen stoplicht fietsen *[breath of 966 ms]* en ik zag een rode auto voor het stoplicht staan [42; 366; 792] en uh op het moment dat zij [40; 374; 896] uh voor de auto langs bijna reed begon de rode auto te rijden ik denk dus dat hij door rood reed. uh I saw a [40; 364; 804] woman on the bike at a uh traffic light [54; 352; 910] pass a green traffic light *[breath of 966 ms]* and I saw a red car standing in front of the traffic light [42; 366; 792] and uh at the very moment that she [40; 374; 896] uh almost cycled past in front of the car the red car started to drive so I think that his light was red.
Non-native speech fragment
uh ik z ik heb gezien dat dat die vrouw was aan het [136; 467; 905] rijden [120; 466; 939] toen uh met een groene licht op de fiets en een auto kwam van die uh rechterkant uh was een rooie auto *[breath of 1001 ms]* die man heeft uh tegen die vrouw [143; 481; 913] gereden [137; 482; 955] en uh [138; 474; 907] ja ik heb de wel een uh rode licht denk ik want die uh die van die vrouw was nog uh groen. uh I z I have seen that that woman was [136; 467; 905] driving [120; 466; 939] when uh with a green light on the bike and a car came from the uh right side uh was a red car *[breath of 1001 ms]* that man has uh against the woman [143; 481; 913] driven [137; 482; 955] and uh [138; 474; 907] yeah I have the well a uh red light I think because that uh that of that woman was still uh green.

Prior to running the rating experiment, all items were evaluated for naturalness in a blinded control procedure by the first author. If a particular manipulated silent pause was perceived as unnatural, its duration was slightly altered while maintaining the range of silent pause durations of each manipulation condition. After the first corrections, the evaluation procedure was repeated by the last author. Finally, the second author listened to all the items and again corrections were made. If specific manipulated pauses were still deemed to sound unnatural after all these corrections, this particular pause was exempted from manipulation in all conditions. Table 3.3 summarizes the differences between the three conditions of Experiment 1 for both native and non-native speech. All resulting audio stimuli were scaled to an intensity of 70dB.

Table 3.3: Pause characteristics of native and non-native speech in the three conditions of Experiment 1 ($N = 60$ per column; M (SD)).

		NoPauses	ShortPauses	LongPauses
Native	Number of pauses per 100 syllables	0 (0)	6.1 (2.0)	6.1 (2.0)
	Silent pause duration (ms)	0 (0)	383 (40)	867 (32)
Non-native	Number of pauses per 100 syllables	0 (0)	6.5 (2.2)	6.5 (2.2)
	Silent pause duration (ms)	0 (0)	393 (32)	873 (29)

Note. Silent pause threshold 150 ms.

Procedure The manipulated versions of the speech fragments (i.e., no original recordings) were presented to participants by making use of the FEP experiment software (Veenker, 2006). Each experimental session started with written instructions, presented on the screen, which instructed participants to judge the speech fragments for overall fluency. Participants were instructed *not* to rate the items in a broad interpretation of fluency (i.e., overall language proficiency, as in: "he is fluent in French"). In contrast, the raters were asked to base their judgments on the use of silent and filled pauses, the speed of delivery of the speech and the use of hesitations and/or corrections (see 6.5). The findings from Chapter 2 of this dissertation have demonstrated that raters, given these instructions, are able to give fluency ratings that correlate strongly with pause and speed measures. Pinget et al. (in press) reported that fluency ratings of this type are relatively independent from such interfering factors as perceived accent. The participants rated the speech fragments using an Equal Appearing Interval Scale (EAIS; Thurstone, 1928): it included nine stars with labelled extremes ('not fluent at all' on the left; 'very fluent' on the right).

Following these instructions but prior to the actual rating experiment four

practice items were presented so that participants could familiarize themselves with the task and the items. The participants were given the opportunity to ask questions if they thought they did not understand the task. No instructions other than the written instructions were supplied to the participants by the experimenters.

After the practice items, the experimental session started. Participants listened to the speech fragments over headphones at a comfortable volume in sound-attenuated booths. The experimental items were arranged in a Latin Square design: participants heard each item in only one condition, with three groups of listeners for counterbalancing. Participants themselves were unaware of this partitioning. In line with the three listener groups, there were three different pseudo-randomised presentation lists of the stimuli and three reversed versions of these lists resulting in six different orders of items.

Each session lasted approximately 45 minutes, but participants were allowed to take a brief pause halfway through the experiment. As introduced previously, at the end of each session the participant filled out a short questionnaire which inquired about personal details, prior experiences with teaching L2 Dutch and/or rating fluency and which factors they thought had influenced their judgments. We also inquired whether they had noticed anything particular about the speech stimuli (as explained under *Participants*).

3.2.2 Results of Experiment 1

Cronbach's alpha coefficients, as measures of interrater agreement, were calculated using the ratings within the three participant groups ($\alpha_1 = 0.95$; $\alpha_2 = 0.96$; $\alpha_3 = 0.95$). Linear Mixed Models (Baayen et al., 2008; Lachaud & Renaud, 2011; Quené & Van den Bergh, 2004, 2008) as implemented in the `lme4` library (Bates et al., 2012) in R (R Development Core Team, 2012) were used to analyze the data (see Appendix B for a link to the raw data).

Our analyses consisted of two phases. In the first phase a correction procedure was carried out. A model was built with random effects for individual differences between speakers (Speaker), individual differences between raters (Rater) and individual differences in order effects, varying within raters (Order). Simple models, containing one or two of these predictors, were compared to more complex models that contained one additional predictor. In order to allow such comparisons of models in our analysis, coefficients of models were estimated using the full maximum likelihood criterion (Hox, 2010; Pinheiro & Bates, 2000). Likelihood ratio tests (Pinheiro & Bates, 2000) showed that the most complex model proved to fit the data of Experiment 1 better than any simpler model. This model showed effects of Speaker $u_{0(j0)}$, Rater $v_{0(0k)}$ and Order, varying within raters, $w_{Order0(0k)}$ and contained a residual component $e_{i(jk)}$. Extending this model with a fixed effect Order, testing for general

learning or fatigue effects, did not improve it ($\chi^2(1) < 1$). Furthermore, we also tested a supplementary model with a maximal random part including random slopes (cf. Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). Because this did not lead to a different interpretation of results, we only report the model with a simple random part.

The second phase of our analyses involved the addition of fixed effects to the model. These fixed effects tested for effects of our particular interest, resulting in the model given in Table 3.4. A fixed effect of Nativeness γ_A was included to test for differences between native and non-native speakers. In the contrasts matrix, native speech was coded with 0.5 and non-native speech with -0.5. Two Condition contrasts were tested: the first contrast γ_B compared the NoPauses condition (contrast coding -0.5) against the ShortPauses and LongPauses conditions (each receiving the contrast coding of 0.25), thus testing for an effect of the *number* of silent pauses. The second contrast γ_C compared the ShortPauses condition (-0.5) against the LongPauses condition (0.5), thus testing for an effect of the *duration* of silent pauses. Matching our first research question, interactions between the two Condition contrasts and the factor Nativeness were also included (γ_D and γ_E), thus testing whether the effect of the number or the duration of silent pauses differed across native and non-native speakers. Finally, fixed effects of the topics tested for differences between the three speaker topics (denoted as γ_F and γ_G). Adding additional interactions between fixed effects did not improve the model: neither interactions between topics and the two Condition contrasts ($\chi^2(4) = 7.0688, p = 0.1323$) nor three-way interactions between topics, Nativeness and the two Condition contrasts ($\chi^2(8) = 14.603, p = 0.06734$) significantly improved the predictive power of the model. No effect of the L1 background of our non-native speakers (Turkish vs. English) was observed and, therefore, this factor was excluded from analysis. The additional interaction between Nativeness and Topic (γ_H and γ_I) did improve the model and was therefore included. Results of this model are listed in Table 3.4. Degrees of freedom (df) required for statistical significance testing of t values was given by $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 20$ speakers) and m is the total number of explanatory variables in the model ($m = 13$) resulting in $df = 6$. In Figure 3.1 the mean fluency ratings are represented graphically.

The significant effect of Nativeness showed that native speakers were rated as more fluent than non-native speakers. Also, both condition contrasts were found to be statistically significant: the condition NoPauses was rated as more fluent than the conditions LongPauses and ShortPauses taken together (the number contrast γ_B), and the condition ShortPauses was rated as more fluent than the LongPauses condition (the duration contrast γ_C). The effects of the manipulations on fluency ratings did not differ between native and non-native speakers, that is, no interaction between either of the two condition contrasts

Figure 3.1: Mean fluency ratings in Experiment 1 (error bars enclose 1.96 × SE, 95% CIs). Plot points were jittered along the x-axis to avoid overlap of error bars.

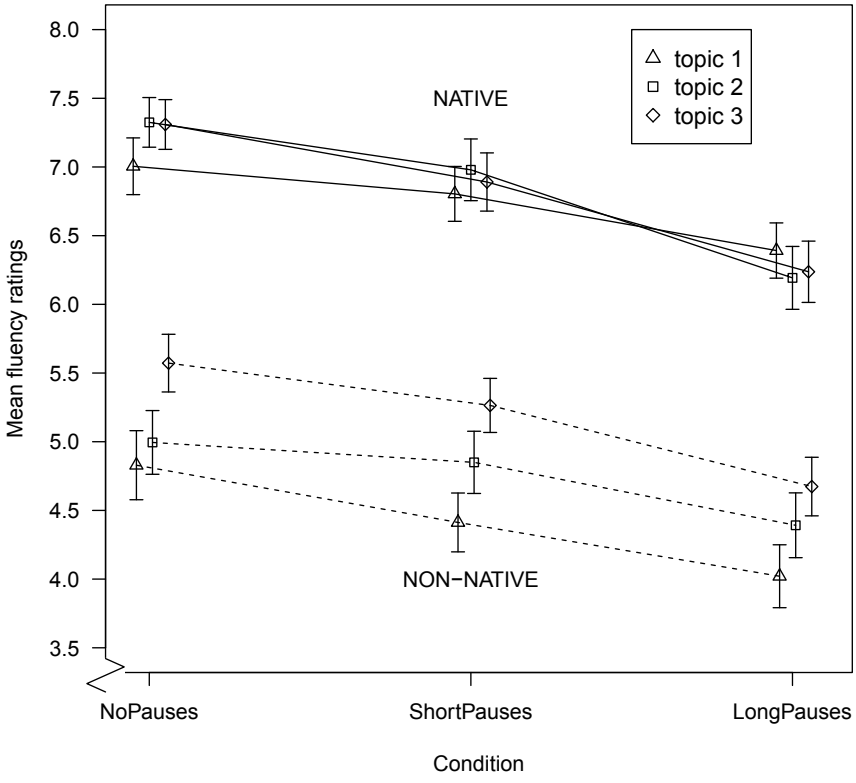


Table 3.4: Estimated parameters of mixed-effects modelling on Experiment 1 (standard errors in parentheses).

	estimates	t values	significance (df=6)
fixed effects			
Intercept, $\gamma_{0(00)}$	5.58 (0.15)	37.75	$p < 0.001$ ***
Nativeness, $\gamma_{A(00)}$	2.33 (0.24)	9.84	$p < 0.001$ ***
Number contrast, $\gamma_{B(00)}$	-0.79 (0.06)	-13.06	$p < 0.001$ ***
Duration contrast, $\gamma_{C(00)}$	-0.55 (0.05)	-10.45	$p < 0.001$ ***
Nativeness x Number contrast, $\gamma_{D(00)}$	-0.18 (0.12)	-1.45	$p = 0.197$
Nativeness x Duration contrast, $\gamma_{E(00)}$	-0.17 (0.11)	-1.62	$p = 0.156$
Topic 2, $\gamma_{F(00)}$	0.21 (0.05)	3.96	$p = 0.007$ **
Topic 3, $\gamma_{G(00)}$	0.42 (0.05)	7.96	$p < 0.001$ ***
Nativeness x Topic 2, $\gamma_{H(00)}$	-0.25 (0.11)	-2.4	$p = 0.053$
Nativeness x Topic 3, $\gamma_{I(00)}$	-0.70 (0.11)	-6.63	$p < 0.001$ ***
random effects			
Speaker intercept, $\sigma^2_{u_{0(j0)}}$	0.25		
Rater intercept, $\sigma^2_{v_{0(0k)}}$	0.46		
Order, $\sigma^2_{w_{Order0(0k)}}$	$< .01$		
Residual, $\sigma^2_{e_{i(jk)}}$	1.59		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

and Nativeness was found. However, effects of the different topics were found in non-native speech: the significant interaction between Topic 3 and Nativeness showed that only non-native speech fragments on topic 3 were rated to be more fluent as compared to topic 1. It is possible to estimate how much of the variability of the fluency ratings the model accounts for by calculating the proportional reduction in unexplained variance (Snijders & Bosker, 1999, p. 99-103). The proportion of explained variance was estimated by comparing the random variance of the full model (in Table 3.4) to the simple model without fixed effects. The proportional reduction in unexplained variance of the full model relative to simple model was 0.343. We also investigated what proportion of the predicted error was accounted for by our manipulation conditions (the Number and the Duration contrasts). For this we compared the full model with a simpler model without the Number and Duration contrasts as predictors. The proportional reduction in unexplained variance was then found to be 0.055. This means that our manipulations accounted for 5.5% of the predicted error.

In Experiment 1, one interaction involving the factor Nativeness was found, namely the interaction between Topic and Nativeness. Our models showed that non-natives were rated as more fluent when talking about topic 2 and 3 than when talking about topic 1 (cf. Table 3.1), but this effect was absent in na-

tive speech. There may have been acoustic differences between the topics in non-native speech. For instance, compared to natives, non-natives could have produced more filled pauses when talking about topic 1 relative to topics 2 and 3. This was assessed in post-test 1 in which the acoustic differences between topics in native and non-native speech were investigated using Linear Mixed Models.

Based on transcriptions of the speech stimuli, acoustic speech measures were calculated for the stimuli in all three manipulation conditions. The speech measures that were investigated were: i) the number of silent pauses per second spoken time, ii) the number of filled pauses per second spoken time, iii) the log of the mean silent pause duration, iv) the log of the mean syllable length, v) the number of repetitions per second spoken time and vi) the number of corrections per second spoken time. We tested models that predicted these acoustic speech measures using fixed effects of Topic, Nativeness and Condition and their interaction (and the random effect Speaker). Indeed one interaction between Topic, Nativeness and Condition was found: non-natives produced significantly fewer silent pauses when talking about topic 3 relative to topic 1 (in the two conditions in which silent pauses were present, namely, ShortPauses and LongPauses). Thus discussing a more difficult topic pushed the non-native speakers in our sample to speak more fluently. The decrease in the production of silent pauses may explain, at least in part, the higher ratings of non-native speech from topic 3.

Another possible account for why non-natives were rated to be more fluent when talking about topics 2 and 3 may possibly be found in the relative difficulty of the topics. Hulstijn et al. (2012) established that successfully produced speech on topic 3 would demonstrate a higher CEFR language proficiency level (B2) than speech on topic 1 or 2 (B1). Adopting this classification of the speaking tasks, raters may have considered the possibly more elaborate vocabulary of the topic when judging fluency. This was investigated in post-test 2 which analysed the frequency of occurrence of the words produced by native and non-native speakers. To test whether more complex speaker topics lead to more complex language among non-native speakers, vocabulary differences between topics were investigated in post-test 2 using Linear Mixed Models. The frequency of occurrence of each token in our speech materials was obtained from SUBTLEX-NL, a database of Dutch word frequencies based on 44 million words from film and television subtitles (Keuleers, Brysbaert, & New, 2010). We tested models that predicted the log frequency of each token using Topic and Nativeness and their interaction as fixed effects and Speaker as random effect. One interaction between Topic and Nativeness was found: non-natives produced more low-frequency words in fragments from topic 3 relative to topic 1, whereas this did not apply to natives. Thus discussing a more difficult topic pushed non-natives to use more low-frequency words. Listeners may have been

influenced by lexical sophistication in their assessment of the complexity of the different topics, which may have caused the higher ratings of non-native speech from topic 3.

3.2.3 Discussion of Experiment 1

In summary, Experiment 1 was designed to provide an answer to the question of how listeners weigh the fluency characteristics of native and non-native speech. Therefore, Experiment 1 targeted the effect of the number of silent pauses and the effect of the duration of silent pauses on both native and non-native fluency perception. Native and non-native speech was manipulated such that there were three experimental conditions: NoPauses (<150 ms), ShortPauses (250-500 ms) and LongPauses (750-1000 ms). Participants who reported to have noticed pause manipulations in the speech stimuli were excluded from the analyses ($n = 14$). Adding these participants to the analyses did not lead to a different interpretation of results.

The high Cronbach's alpha coefficients demonstrated that the raters strongly agreed amongst each other. The main effect of Nativeness showed that, overall, natives were perceived to be more fluent than non-natives (a difference of 2.33 on our 9-point scale). The native and non-native speech had been matched for the number of silent pauses, but still differed in other aspects which have been shown to contribute to fluency perception (Cucchiarini et al., 2002; Ginther et al., 2010; Rossiter, 2009): non-natives produced more filled pauses (*uh*) per second spoken time, more repetitions per second spoken time, and had longer syllable durations than natives. Any of these temporal but also non-temporal factors (e.g., vocabulary, grammar, etc.) may have contributed to the fact that, overall, non-native speech was rated to be less fluent than native speech.

Furthermore, it has been observed that pauses in native speech occur in different positions in the sentence as compared to those in non-native speech (e.g., Skehan & Foster, 2007). Our native and non-native speech materials had been matched for silent pauses, but pause distribution was not taken into account. If pauses in our native materials occurred in different positions than those in our non-native materials, it may be expected that there would be differential effects of our manipulation conditions across native and non-native speech. However, inspection of our stimuli showed that our speech fragments of approximately 20 seconds were too short to provide the listener with a firm idea of pause distribution (number of pauses in between AS-units per speech fragment: $M_{L1} = 1.5$; $M_{L2} = 1$).

It was also established that increasing the number of silent pauses whilst keeping all other possibly interacting factors constant, led to a decrease in fluency ratings. More specifically, the addition of one pause every 15 syllables (approximately; see Table 3.3) led to an average decrease in fluency ratings

of 0.79 on the 9-point scale. Also, increasing the duration of silent pauses resulted in a decrease in fluency judgments: lengthening pauses by roughly 480 ms (see Table 3.3) led to an average decrease in fluency ratings of 0.55 on our 9-point scale. These effects, together with the proportional reduction in unexplained variance of 0.055, may seem to be relatively small contributions of silent pauses to fluency judgments. However, one should note that silent pauses are not the only contributors to perceived fluency ratings. The observed variance in perceived fluency may be explained by a range of factors, such as silent pauses but also filled pauses, speaking rate, corrections, repetitions, etc. As such, our results are in line with previous research (e.g., Cucchiarini et al., 2002; Ginther et al., 2010), showing that both the *number* and the *duration* of silent pauses have significant effects on fluency ratings. The approach of the current study (manipulating speech in one factor whilst keeping all else constant) has allowed us i) to attribute the observed effects to controlled manipulated variables, and ii) to distinguish between the contributions of the two properties of silent pauses.

With respect to the two hypotheses mentioned earlier, our statistical model did not show any difference in the effects of our manipulations across native and non-native speech. There was no indication that the manipulations affected native speech any differently from non-native speech. Natives were rated more fluent than non-natives, and manipulations of silent pauses led to lower fluency ratings, with no discernible differences between native and non-native speech.

Two post-tests were run to investigate the observed interaction between Topic and Nativeness. These post-tests demonstrated that acoustic differences between topics in non-native speech, and the vocabulary of the non-native speech from topic 3 may have influenced raters in Experiment 1 to rate non-natives to be more fluent when talking about topic 2 and 3 relative to topic 1. Still other factors that we did not control for and have not investigated further can be argued to have influenced the raters (e.g., grammatical accuracy). All these differences between native and non-native speech may have been partially responsible for the difference between natives and non-native speech in fluency perception. However, these differences between natives and non-natives were independent from our experimental manipulations. We found no indications for differential effects our pause manipulations on the perception of fluency in native versus non-native speech.

3.3 Experiment 2

In addition to the speaker's pausing behavior, the speed of speech has been shown to play an important role in fluency perception (cf. Cucchiarini et al., 2002, and the findings from Chapter 2 of this dissertation). Experiment 2 ex-

tends the insights from Experiment 1 by studying the effect of the speed of speech on fluency ratings of native and non-native speech. The original native and non-native speech materials from Experiment 1 (i.e., not the manipulated versions) were re-used and manipulated in terms of the speed with which speakers were speaking.

Previously, Munro and Derwing (1998, 2001) also applied speed manipulations to native and non-native speech. Munro and Derwing (1998), in their Experiment 2, adjusted the speaking rate of native English speech to the mean speaking rate of L2 English speakers and vice versa. Their dependent variable was the rated 'appropriateness of the speed'. They found that some speeded non-native speech was found to be more 'appropriate' than unmodified non-native speech. Munro and Derwing (2001) made use of speed manipulations to study different dependent variables, namely perceived accentedness and comprehensibility. In that study, only non-native speech materials were analyzed. Results indicated that the speaking rate could account for 15% of the variance in accentedness ratings. The phonetic manipulations in both studies by Munro and Derwing involved speech compression-expansion applied to the entire speech signal including silences. This entails that not only the articulation rate but also the duration of the pauses was altered (i.e., manipulations of speech rate).

In the present Experiment 2, the dependent variable is perceived fluency. Because in the materials of Experiment 1 the articulation rate of the native speakers was not matched to the articulation rate of the non-native speakers (see the discussion of Experiment 1 above), we used a cross-wise experimental design to match the two groups (cf. Munro & Derwing, 1998). The speed of non-native speech was sped up to the mean value of the native speakers, and the native speech was slowed down to the mean value of non-native speakers. This procedure made comparisons across native and non-native speakers possible. The increase in speed in non-native speech is expected to lead to an increase in fluency ratings and the decrease in speed in native speech to a decrease in perceived fluency. The magnitude of these two effects may either be similar or different from each other (e.g., speed manipulations affecting non-native fluency perception more than native fluency perception, or vice versa).

An important distinction between Munro and Derwing's studies and the current work is that not only the *speech rate* (including pauses) but also the *articulation rate* (excluding pauses) was manipulated. Thus the contribution of silent pauses to fluency perception (Experiment 1) was clearly separated from the contribution of the speed of the speech (Experiment 2). Experiment 2 thus consisted of three conditions: the original speech, speech with its speech intervals manipulated (i.e., articulation rate manipulations) and speech with both its speech intervals and its silent intervals manipulated simultaneously (i.e., speech rate manipulations). The effect of manipulations in speech rate are

expected to be larger than manipulations in articulation rate because pause duration has already been shown to contribute to perceived fluency in Experiment 1.

3.3.1 Method of Experiment 2

Participants Seventy-three members of the same UiL OTS participant pool took part in the experiment with implicit informed consent. All were native Dutch speakers with normal hearing (M_{age} = 21.22, SD_{age} = 4.30, 7m/66f). None had previous experience in teaching L2 Dutch or rating fluency. The post-experimental questionnaire inquired (amongst other issues) whether they had noticed anything particular about the experiment. Of all participants, 19 responded that they thought the stimuli had been edited in some way. Again, individual responses ranged from comments about non-native accents to different amplitudes. All responses from participants which could reasonably be interpreted as relevant to the pause and also the speed manipulations were taken as evidence of awareness of the experimental manipulation (n = 11; 15%). Data from these participants were excluded from the analyses. Data from an additional four participants were lost due to technical reasons. One participant had already taken part in Experiment 1 and, for that reason, was excluded from further analyses. The final dataset included the remaining 57 participants (M_{age} = 21.44, SD_{age} = 3.18, 6m/51f).

Stimulus description The original recordings from the native and non-native speakers from Experiment 1 (i.e., not the pause-manipulated speech fragments) served as the basis of the materials of Experiment 2. As explained above, non-native speech was increased in speed to match the mean speaking rate of the natives and native speech was slowed down to match the mean speaking rate of the non-natives, thus making comparisons across native and non-native speakers possible. Two types of speed manipulations were performed in Experiment 2, relating to two different measures of the speed of speech. Based on manual transcriptions of the speech stimuli, both the speech rate and the articulation rate of every speech fragment was calculated. *Speech rate* is calculated as the number of produced syllables per second of the total time (i.e., including silences). In contrast, *articulation rate* is calculated per second of the spoken time (i.e., excluding silences). In line with this distinction, two types of speed alterations were part of Experiment 2: a manipulation of spoken time and a manipulation of the total time.

Together with the original recording this resulted in three conditions: Original, Articulation Rate Manipulations (ARM) and Speech Rate Manipulations (SRM). In the ARM condition, native speakers were slowed down to the mean value of the non-native speakers (ratio=1.206) and the speed of non-native

speech was increased to the mean value of the native speakers (ratio=0.829). This manipulation was performed only on the speech intervals in between pauses of >250 ms using PSOLA, a method for manipulating the pitch and duration of speech (Pitch-Synchronous OverLap-and-Add; Moulines & Charpentier, 1990) as implemented in Praat (Boersma & Weenink, 2012). The settings used for the manipulation were: minimum frequency=75Hz, maximum frequency females=420Hz, maximum frequency males=220Hz. In this manner, items in the ARM condition differed from the original speech only in the speed of articulation. The duration of silent pauses was identical in both conditions. Table 3.5 provides examples exemplifying the three manipulation conditions.

Table 3.5: Examples of speech fragments on topic 1 from a native and non-native speaker. Durations of speech intervals (ms) are given in bold as {**Original; ARM; SRM**} and subsequently silent pause durations as [Original; ARM; SRM]. Translations can be found in Table 3.2.

Native speech fragment
uh ik zag een {**1150; 1387; 1387**} [562; 562; 655] vrouw op de fiets bij een uh stoplicht {**3382; 4080; 4080**} [341; 341; 397] door een groen stoplicht fietsen {**1772; 2138; 2138**} *[breath of 966 ms]* en ik zag een rode auto voor het stoplicht staan {**3105; 3746; 3746**} [609; 609; 710] en uh op het moment dat zij {**1986; 2397; 2397**} [349; 349; 407] uh voor de auto langs bijna reed begon de rode auto te rijden ik denk dus dat hij door rood reed {**7622; 9085; 9085**}
Non-native speech fragment
uh ik z ik heb gezien dat dat die vrouw was aan het {**2535; 2102; 2102**} [433; 433; 359] rijden {**520; 431; 431**} [373; 373; 308] toen uh met een groene licht op de fiets en een auto kwam van die uh rechterkant uh was een rooie auto {**6905; 5723; 5723**} *[breath of 1001 ms]* die man heeft uh tegen die vrouw {**2028; 1681; 1681**} [545; 545; 452] gereden {**883; 732; 732**} [835; 835; 692] en uh {**792; 657; 657**} [1209; 1209; 1002] ja ik heb de wel een uh rode licht denk ik want die uh die van die vrouw was nog uh groen {**5648; 4682; 4682**}

Native speech fragments that had an exceptionally slow articulation rate (such that, after manipulation, they would fall below the slowest speaking rate of the non-native speakers) were either, prior to the standard manipulation, changed to non-outlier value ($n = 3$), or they were slowed down with a smaller ratio (i.e., a ratio of 1.166; $n = 1$) such that it matched the syllable duration of the slowest non-native speech fragment. A similar procedure was adopted for exceptionally fast non-native speech fragments: they were either changed to non-outlier value ($n = 2$) or their speed was increased with smaller ratios (0.877 and 0.904; $n = 2$) such that they matched the syllable duration of the fastest native speech fragment. Similar to the method of Experiment 1, all manipulated items were evaluated for their naturalness by the first author, and

corrected accordingly. Subsequently, this procedure was repeated by the last and, finally, also by the second author. For instance, four very fast non-native sentences within the speech fragments and seven very slow native sentences were exempted from manipulation.

In the SRM condition, the same modifications in native and non-native speech were made as in the ARM condition but this time the manipulation was performed on the entire speech fragment including the silent pauses. Thus, items in the SRM condition differed from the ARM condition only in the duration of silent pauses. The speed of articulation was identical in the ARM and SRM condition. Table 3.6 summarizes the differences between conditions of Experiment 2 for both native and non-native speech. This table illustrates that the values for the two manipulation conditions of native speech were matched to the original values of non-native speech (and vice versa). All resulting audio stimuli were scaled to an intensity of 70dB.

Table 3.6: Speed characteristics of native and non-native speech in the three conditions of Experiment 2 ($N = 60$ per column; M (SD), [Range]).

		Number of syllables per second spoken time (articulation rate)	Number of syllables per second total time (speech rate)
Native	Original	4.87 (0.53), [3.86-5.72]	3.94 (0.51), [3.26-5.13]
	ARM	4.04 (0.44), [3.20-4.74]	3.37 (0.41), [2.77-4.33]
	SRM	4.04 (0.44), [3.20-4.74]	3.26 (0.42), [2.70-4.26]
Non-native	Original	3.88 (0.39), [3.20-4.79]	3.26 (0.42), [2.41-4.37]
	ARM	4.68 (0.47), [3.86-5.78]	3.82 (0.53), [2.77-5.17]
	SRM	4.68 (0.47), [3.86-5.78]	3.94 (0.51), [2.91-5.27]

Note. Silent pause threshold 250 ms.

Procedure The pseudo-randomization, post-experimental questionnaire, instructions, and scales in Experiment 2 were the same as those used in Experiment 1.

3.3.2 Results of Experiment 2

Cronbach's alpha coefficients were calculated on the ratings within the three participant groups ($\alpha_1 = 0.93$; $\alpha_2 = 0.93$; $\alpha_3 = 0.92$). Similar to the analyses in Experiment 1, the ratings were analyzed using Linear Mixed Models (see Appendix B for a link to the raw data). Again, random effects of Speaker, Rater and Order, varying within raters, were included in the model. We also tested a supplementary model with a maximal random part including random slopes (cf. Barr, 2013; Barr et al., 2013). Because this did not lead to a different

interpretation of results, we only report the model with a simple random part. Subsequently, fixed effects were added to the model, resulting in the model given in Table 3.7.

Similar to the model of Experiment 1, a fixed effect of Nativeness (γ_A) compared ratings of native items with ratings of non-native items. Again, native speech was coded with 0.5 and non-native speech with -0.5. A fixed effect of ARM (γ_B) tested for differences between original versions and ARM versions. In the contrast matrix, the original speech received the coding -0.5 and the manipulated speech the code 0.5. Also an interaction with Nativeness was included (γ_C). Recall that the articulation rate was manipulated in two directions: the articulation rate in non-native speech was increased whereas it was slowed down in native speech. If the speed manipulations would affect native speech to a similar extent as non-native speech, then it is expected that slowed down native speech would lead to a decrease in fluency ratings, and that non-native speech that has been increased in speed would lead to an increase in fluency ratings. In a statistical analysis the decrease in native fluency and the increase in non-native fluency are, then, expected to cancel each other out. Therefore, we do not expect to find a main ARM effect (γ_B) but rather an interaction with Nativeness (γ_C). However, if the speed manipulations affect native speech differently from non-native speech, this would have to show in a main effect of ARM (γ_B). The same holds for the SRM condition; a fixed main effect of SRM and an interaction with Nativeness (γ_D and γ_E) were also included.

In addition, a fixed effect of Topic (γ_F and γ_G) was included to investigate main topic effects, along with interactions between Topic and Nativeness (γ_H and γ_I). A fixed effect of Order (γ_J), testing for overall learning or fatigue effects, improved the explanatory power of the model and was therefore included in the model. No effect of the L1 background of our non-native speakers (Turkish vs. English) was observed and, therefore, this factor was excluded from the analysis.

The estimates from our statistical model are listed in Table 3.7. Degrees of freedom required for testing of statistical significance of t values was computed as follows: $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 20$ speakers) and m is the total number of explanatory variables in the model ($m = 14$) resulting in $df = 5$. Figure 3.2 illustrates mean uency ratings from this experiment.

A significant effect of Nativeness showed that, overall, native speakers were rated as more fluent than non-native speakers. With respect to the ARM condition, no main effect of ARM was found but only an interaction with Nativeness. This interaction reflected the different directions of the ARM manipulations. Slowed down native speech was rated as less fluent than the original native speech, and non-native speech that had received an increased speed was rated as more fluent than the original non-native speech. The decrease in fluency

Figure 3.2: Mean fluency ratings in Experiment 2 (error bars enclose 1.96 × SE, 95% CIs). Plot points were jittered along the x-axis to avoid overlap of error bars.

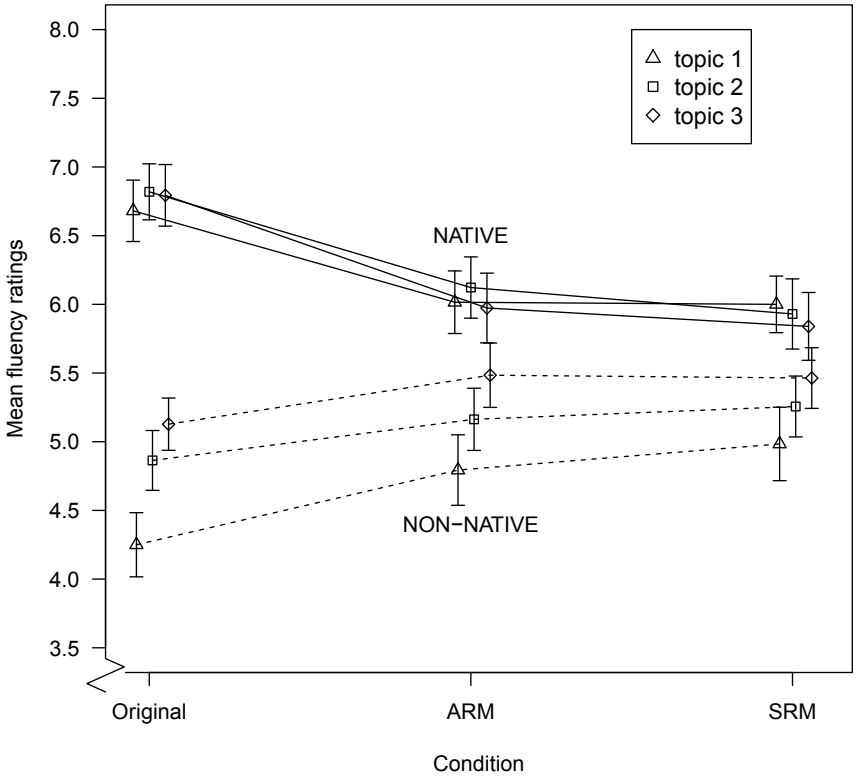


Table 3.7: Estimated parameters of mixed-effects modelling on Experiment 2 (standard errors in parentheses).

	estimates	t values	significance (df=5)
fixed effects			
Intercept, $\gamma_{0(00)}$	5.45 (0.17)	32.31	$p < 0.001$ ***
Nativeness, $\gamma_{A(00)}$	1.57 (0.29)	5.32	$p = 0.003$ **
ARM, $\gamma_{B(00)}$	-0.09 (0.06)	-1.38	$p = 0.226$
ARM x Nativeness, $\gamma_{C(00)}$	-0.64 (0.13)	-4.84	$p = 0.005$ **
SRM, $\gamma_{D(00)}$	-0.14 (0.06)	-2.11	$p = 0.089$
SRM x Nativeness, $\gamma_{E(00)}$	-1.11 (0.13)	-8.41	$p < 0.001$ ***
Topic 2, $\gamma_{F(00)}$	0.24 (0.06)	4.22	$p = 0.008$ **
Topic 3, $\gamma_{G(00)}$	0.33 (0.06)	5.82	$p = 0.002$ **
Nativeness x Topic 2, $\gamma_{H(00)}$	-0.38 (0.11)	-3.38	$p = 0.019$ *
Nativeness x Topic 3, $\gamma_{I(00)}$	-0.73 (0.11)	-6.48	$p = 0.001$ **
Order, $\gamma_{J(00)}$	-0.01 (0.00)	-2.50	$p = 0.054$
random effects			
Speaker intercept, $\sigma^2_{u_{0(j0)}}$	0.40		
Rater intercept, $\sigma^2_{v_{0(0k)}}$	0.39		
Order, $\sigma^2_{w_{Order0(0k)}}$	$< .01$		
Residual, $\sigma^2_{e_{i(jk)}}$	1.78		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

perception in native speech was found to be similar to the increase in perceived fluency in non-native speech, as evidenced by the absence of a main effect of ARM. A similar picture is observed for the SRM condition: no effect of this condition was found, but the interaction with Nativeness was statistically significant. The effect of the SRM manipulation was, as expected, larger than the effect of the ARM manipulation (i.e., the effect of SRM x Nativeness was larger than the effect of ARM x Nativeness). In addition, main effects of Topic were found and also interactions with Nativeness, namely, in non-native speech, the more difficult topics (2-3) were rated higher in fluency than the easy topic (1). Finally, also a very small, statistically marginal overall order effect was found. The proportion of explained variance was estimated through a comparison of the random variance of the full model, given in Table 3.7, and the simple model without any fixed effects: 0.161. The proportional reduction in unexplained variance that was due to the manipulation conditions (i.e., the ARM and the SRM predictors) was estimated by comparing the full model to a simpler model without ARM and SRM as predictors. The proportional reduction in unexplained variance was then found to be 0.035. This means that our manipulations accounted for 3.5% of the predicted error.

3.3.3 Discussion of Experiment 2

In summary, Experiment 2 was designed to provide an answer to the question of how listeners weigh the fluency characteristics of native and non-native speech. Therefore, Experiment 2 focused on the effect of the speed of the speech on both native and non-native fluency perception. Native and non-native speech was manipulated such that there were three conditions: original recordings, recordings that had been manipulated in their articulation rate (ARM) and recordings that had been manipulated in their speech rate (SRM). In these last two manipulated conditions, the direction of the manipulation differed for native and non-native speech: non-native speech was increased to match the native speech whereas native speech was slowed down to match the non-native speech. Again, those participants who reported to have noticed the manipulations in the speech stimuli were excluded from the analyses ($n = 11$). Adding these participants to the analyses did not lead to a different interpretation of results.

Statistical analyses demonstrated that, overall, natives were perceived to be more fluent than non-natives (a difference of 1.57 on our 9-point scale). This effect replicates the Nativeness effect found in Experiment 1. It was expected that the increase in speed in non-native speech would lead to an increase in fluency ratings and that the decrease in speed in native speech would lead to a decrease in perceived fluency. The statistical analyses corroborated this expectation. Crucially, the relative increase and decrease in fluency ratings were of similar magnitude. Natives were rated higher than non-natives overall, with no indication that manipulation in the speed of speech affected natives and non-natives differently. Similar to Experiment 1, an interaction between Topic and Nativeness was found: non-natives were rated to be more fluent when talking about topic 2 and 3 relative to topic 1 (cf. Table 3.1). Since the same speech materials were used for Experiment 1 and 2, vocabulary differences and acoustic differences between the speech of natives and non-natives may explain this interaction in the same way as for Experiment 1.

The manipulations of speech intervals in between silent pauses (ARM condition) may not only have affected the perception of these speech intervals but also the perception of the duration of the (unedited) silent pauses. Slowing down speech may cause the duration of pauses to be perceived as subjectively shorter. The expected negative effect of slowing down speech on perceived fluency could then be countered by a positive effect of shorter pauses. Although we cannot rule out such a counter-effect in Experiment 2, it certainly was not strong enough to neutralize the primary effect of our speed manipulations. However, we did observe a stronger effect of the manipulation in speech rate (SRM) as compared to the manipulation in articulation rate (ARM), since the former included pauses. In fact, the SRM manipulations can be viewed as a combination

of Experiment 1 (silent pauses) and the ARM manipulation within Experiment 2 (speed): the faster the articulation rate and the shorter the pauses, the higher the fluency ratings, both in native and non-native speech.

3.4 General discussion

The current study carries several implications. First of all, it has demonstrated that fluency characteristics present in the speech signal affect the perception of fluency in both native and non-native speech: the more disfluency in the utterance, the lower the fluency ratings. This observation extends our current knowledge of the concept of fluency. Previous work has shown that such temporal factors as acoustic measures of the speech signal could explain variation in fluency ratings to a large degree (e.g., Cucchiarini et al., 2000, and the findings from Chapter 2 of this dissertation). Non-temporal factors such as perceived foreign accent have been shown to play a much smaller role (e.g., Pinget et al., in press). The finding that the perception of fluency depends on the produced fluency characteristics of speech is relevant, because it confirms that variation in fluency judgments between different speakers can be accounted for by quantitative differences.

Furthermore, our study has demonstrated that the relationship between utterance fluency and perceived fluency is similar across native and non-native speech. Manipulations for four phonetic factors (number of silent pauses, their duration, articulation rate, and speech rate) showed similar effects on perceived fluency for native and non-native speakers. This is a striking result considering that native and non-native speech differ in many respects (e.g., prosody, grammar, lexis, pronunciation, etc.). The main effect of the Nativeness factor in both our experiments testifies to this clear distinction: our listeners easily discriminated native and non-native speakers. Nevertheless, our experiments demonstrate that it is possible, through careful phonetic manipulation, to measure how specific acoustic properties contribute to fluency judgments of native and non-native speech, whilst keeping some other possibly interacting factors constant. Thus, we observe that silent pause manipulations (Experiment 1) and speed manipulations (Experiment 2) affected subjective fluency ratings of native and non-native speech to a similar degree.

Our study has demonstrated that there is no difference in the way listeners weigh the fluency characteristics of native and non-native speech. One should note, however, that we provided our fluency raters with particular instructions to judge the pausing, speed, and repair behavior of the native and non-native speakers. Our instructions were formulated in such a way that raters assessed fluency in its *narrow* sense (Lennon, 1990), as one of the components of speaking proficiency. The alternative to this approach would be to have raters assess

fluency without any instructions on what comprises fluency. This alternative approach is expected to result in ratings of fluency in its *broad* sense (Lennon, 1990), as a synonym of overall speaking proficiency.

There were several reasons why the experiments reported above used ratings of fluency in the narrow sense. First of all, this approach is consistent with previous studies of fluency perception that have also used specific fluency instructions (cf. Derwing et al., 2004; Rossiter, 2009). These studies made use of narrow definitions of uency in instructions given to listeners (compared to broad or undefined instructions), precisely because the authors wished to collect reliable ratings of how listeners interpret uency in its narrow sense as one aspect of spoken language. If, by contrast, the interpretation of the concept of uency would be left up to the listener, considerable variability in the subjective ratings is expected to be the result. The findings from earlier literature indicate that instructing listeners to specifically assess fluency in the narrow sense, results in subjective ratings that can be accounted for to a large extent by the temporal characteristics of the speech signal (see our review of relevant literature in the Introduction).

Another reason for instructing raters to assess the narrow sense of fluency is that this approach is compatible with language testing practice. Many language tests (e.g., TOEFL iBT, IELTS, ACTFL, PTE Academic) use speaking rubrics with explicit mention of different aspects of fluency, such as speed of delivery and hesitations. Therefore, the raters for these tests are provided with explicit instructions about how to assess oral fluency. Our conclusions about the similarity of native and non-native fluency perception, based on subjective ratings of the narrow sense of fluency, are therefore directly applicable to language testing practice where similar methods are used.

Although the narrowly-defined fluency definition adopted in this study is fully compatible with existing empirical and assessment literature, it may still be argued that, by instructing raters to evaluate the pause, speed, and repair behavior of speakers, listeners were discouraged to take into account other factors that may influence fluency assessment with respect to potential differences between native and non-native speech. Thus, our finding of no difference in how listeners perceive native and non-native fluency phenomena could be attributed to the specific nature of the instructions given to listeners in making their fluency judgments. However, our results do not suggest that our specific instructions guided listeners to ignore the distinction between native and non-native speech. In fact, we observed a consistent main effect of the Nativeness factor in both our experiments, testifying to listeners' ability to perceive a reliable difference in their rating of fluency in native and non-native speech. Nevertheless, this perceived distinction between native and non-native speech did not affect the way in which listeners weighed native and non-native fluency characteristics for fluency assessment. Therefore, we conclude that the speci-

ficity of our instructions cannot fully explain why our listeners weighed the fluency characteristics of native and non-native speech in a similar fashion.

Our justifications for collecting ratings targeting the narrow sense of fluency do not imply that an alternative approach to fluency perception (i.e., collecting ratings of fluency defined in its broad sense) should not be pursued. In fact, there have been several studies looking into the factors that contribute to perceived oral proficiency. For instance, Kang, Rubin, and Pickering (2010) reported that a combined set of suprasegmental features of non-native speech (e.g., measures of speech rate, pausing, and intonation) accounted for 50% of the variance in overall proficiency ratings. Ginther et al. (2010) found moderate to strong correlations between overall oral proficiency scores and speech rate, speech time ratio, mean length of run, and the number and length of silent pauses. Taken together, these studies suggest that ratings of fluency in its broad sense are also to a great extent determined by temporal characteristics of non-native speech. It however remains to be shown whether native and non-native fluency characteristics are also *weighed* in a similar fashion when it comes to perceived fluency in its broad sense. As yet, the relationship between the perception of fluency in its broad and narrow sense is under-investigated, and so are potential differences between native and non-native fluency. Our present findings can thus be viewed as an initial attempt to fill these particular gaps in our understanding of fluency perception.

The results of our study carry consequences for how we understand the concept of the native speaker. Disfluencies contribute to the perceived fluency level of native speakers in the same way as they affect non-native fluency levels. From the literature on social psychology (Brown et al., 1975; Krauss & Pardo, 2006), we know that listeners assess the speech of others on an everyday basis. People make attributions about speakers' social status, background and even physical properties (Krauss et al., 2002; Krauss & Pardo, 2006). Our results show that individual differences between native speakers in their production of disfluencies carry consequences for listeners' perceptions of a native speaker's fluency level. Thus, the idea that native speakers are generally fluent by default can be called into question. Indeed, our results add to the on-going debate on the notion of the native speaker. For instance, Hulstijn (2011) advocates that a closer look be given to the distinction between native and non-native, suggesting that the distinction may be a gradient rather than a categorical one. Our study provides some support for this statement, in that our experiments show that variation in fluency production affects subjective fluency judgments. We found no reason to believe that listeners make a qualitative distinction between native and non-native speakers in fluency assessment. This view also has implications for language testing practice. The fluency level of non-native speakers is regularly assessed in language tests on the grounds of an idealized native-speaker norm. Our results has show that there is variation in the per-

ceived fluency of native speakers. As a consequence, we conclude that a single ideal native fluency standard does not exist.

Note that our study does not necessarily warrant the conclusion that native and non-native fluency characteristics are perceptually equivalent. Despite our finding that native and non-native fluency characteristics are weighed similarly by listeners, it is likely that the psycholinguistic origins of native and non-native disfluency in production do differ. Non-native disfluency, for instance, is likely to be caused by incomplete linguistic knowledge of, or skills in the non-native language, whereas this is unlikely for native disfluency. These different psycholinguistic origins of disfluency could lead to different functions of native and non-native disfluencies in speech processing. For instance, it has previously been found that native disfluencies may help the listener in word recognition (Corley & Hartsuiker, 2011), in sentence integration (Corley et al., 2007) and in reference resolution (Arnold et al., 2007). Whether or not non-native disfluencies can have similar functions in speech comprehension, is a question that will be addressed in the following two chapters of this dissertation. The current study, which has revealed no essential differences in the way listeners weigh the fluency characteristics of native and non-native speech, can provide a baseline for future investigations into this and similar issues.

CHAPTER 4

Native 'um's elicit anticipation of low-frequency referents, but non-native 'um's do not[1]

4.1 Introduction

Prediction in human communication lies at the core of language production and comprehension. In speech comprehension, listeners habitually predict the content of several levels of linguistic representation based on the perceived semantics, syntax and phonology of the incoming linguistic signal (Kutas, DeLong, & Smith, 2011; Pickering & Garrod, 2007, 2013). This paper contributes to the notion that listeners form linguistic predictions not only based on *what* is said, but also on *how* it is said, and *by whom*. The focus is on two particular performance characteristics of the speech signal, namely disfluency and foreign accent. Our experiments demonstrate that listeners can attribute the presence of disfluency to the speaker having trouble in lexical retrieval, as indicated by anticipation of low-frequency referents following disfluency. Furthermore, listeners are highly flexible in making these predictions. When listening to speech containing a non-native accent, comprehenders modulate their expectations of the linguistic content following disfluencies.

There is a large body of evidence suggesting that people predict the speech

[1] An adapted version of this chapter has been submitted to an international peer-reviewed journal.

of their conversational partner (see Kutas et al., 2011; Pickering & Garrod, 2007, for reviews). Most research has focused on prediction based on semantic (e.g., Altmann & Kamide, 1999), syntactic (e.g., Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004) or phonological properties (e.g., DeLong, Urbach, & Kutas, 2005) of the linguistic input. Other studies have investigated the way performance aspects of the speech signal may affect prediction, such as prediction based on prosody (Dahan, Tanenhaus, & Chambers, 2002; Weber, Grice, & Crocker, 2006). The current paper studies another performance aspect of the speech signal, namely disfluency.

Disfluencies are "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (Fox Tree, 1995), such as silent pauses, filled pauses (e.g., *uh*'s and *uhm*'s), corrections, repetitions, etc. Disfluency is a common feature of spontaneous speech: it is estimated that six in every hundred words are affected by disfluency (Bortfeld et al., 2001; Fox Tree, 1995). Traditionally, it was thought that the mechanisms involved in speech perception are challenged by the disfluent character of spontaneous speech (Martin & Strange, 1968). It was assumed to pose a continuation problem for listeners (Levelt, 1989), who were thought to be required to edit out disfluencies in order to process the remaining linguistic input. Thus, disfluencies would uniformly present obstacles to comprehension and need to be excluded in order to study speech comprehension in its 'purest' form (cf. Brennan & Schober, 2001). However, experimental evidence has shown that disfluencies may help the listener. They may aid comprehenders to avoid erroneous syntactic parsing (Brennan & Schober, 2001; Fox Tree, 2001), to attenuate context-driven expectations about upcoming words (Corley et al., 2007; MacGregor et al., 2010), and to improve recognition memory (Collard et al., 2008; Corley et al., 2007; MacGregor et al., 2010).

Arnold and colleagues have demonstrated that disfluencies may also guide prediction of the linguistic content following the disfluency (Arnold et al., 2003, 2007, 2004). In the two earlier studies, Arnold and colleagues investigated whether listeners use the increased likelihood of speakers to be disfluent (e.g., saying 'thee uh candle' instead of 'the candle') while referring to new as compared to given information (Arnold et al., 2000) as a cue to the information structure of the utterance. In eye-tracking experiments using the Visual World Paradigm, participants' eye fixations revealed that, prior to target onset, listeners were biased to look at a discourse-new referent when presented with a disfluent utterance: a *disfluency bias* toward discourse-new referents. In contrast, when listening to a fluent instruction, listeners were more likely to look at a given object rather than a new object, which is consistent with the general assumption that given information is more accessible than new information. Arnold et al. (2007) extended the disfluency bias to the reference resolution of

known vs. unknown objects (cf. Watanabe et al., 2008). Upon presentation of a disfluent sentence such as 'Click on thee uh red [target]', listeners were found to look more at an unknown object (an unidentifiable symbol) prior to target onset as compared to a known object (e.g., an ice-cream cone).

Additional experiments in Arnold et al. (2007) and Barr and Seyfeddinipur (2010) targeted the cognitive processes responsible for the disfluency bias. In the second experiment reported in Arnold et al. (2007), the authors tested whether (1) listeners 'simply' associated unknown or discourse-new referents with disfluency, or that (2) listeners actively made rapid inferences about the source of the disfluency (e.g., when the speaker is perceived to have trouble in speech production, the most probable source of difficulty is the unfamiliarity of the unknown referent). This second experiment was identical to their first experiment, except that participants were now told that the speaker suffered from object agnosia (a condition involving difficulty recognizing simple objects). Based on this knowledge about the speaker, listeners might predict the speaker to have equal difficulty in naming known and unknown objects, and, therefore, be equally disfluent for both types of targets. Results revealed that the preference for unknown referents following a disfluency, observed in the first experiment, disappeared in the second experiment. This suggests that listeners draw inferences about the speaker's cognitive state which modulates the extent to which disfluency guides prediction.

According to Barr and Seyfeddinipur (2010), the mechanism responsible for the disfluency bias is a perspective-taking process. They investigated whether the disfluency bias for discourse-new referents from Arnold et al. (2003) indicates a preference for referents that are discourse-new for *the listener* or for *the speaker*. By presenting participants with different speakers they could modulate the discourse-status of objects *from the speaker's perspective* while maintaining the discourse-status of the objects *for the listener*. They found that listeners who heard a disfluency directed their attention toward referents that were new for the person speaking, showing that the disfluency bias was dependent on, not just the givenness from the listener's point of view, but on what was old and new for the speaker at hand. The results from Barr and Seyfeddinipur (2010) and Arnold et al. (2007) argue against an egocentric theory of reference resolution (cf. Barr & Keysar, 2006; Keysar, Barr, Balin, & Brauner, 2000; Pickering & Garrod, 2004). Instead, listeners take the speaker's perspective and knowledge into account in real-time speech processing.

Based on the literature, we conclude that (1) listeners are sensitive to disfluencies in the speech signal, (2) disfluencies may direct listeners' expectations in reference resolution, (3) this disfluency bias towards discourse-new or unknown referents involves drawing an inference about the origins of disfluency, and (4) these inferences may be modulated by the listener's model of the assumed speaker's cognitive processes. But what does it mean to draw inferences about

the source of disfluency? The fluency framework described in Segalowitz (2010) provides a theoretical model of how listeners attribute the presence of disfluency to difficulty in speech production. In this framework, adapted from Levelt (1989) and De Bot (1992), the fluency of an utterance is defined by the speaker's *cognitive fluency*: the operation efficiency of speech planning, assembly, integration and execution. Thus, the underlying causes of disfluency may be viewed as corresponding to different stages in speech production. The critical points in speech production where underlying processing difficulty could be associated with speech disfluencies are termed 'fluency vulnerability points' (Segalowitz, 2010, Figure 1.2). For instance, disfluency can originate from trouble in finding out what to say (conceptualization), choosing the right words (formulation), generating a phonetic plan (articulation), or problems in self-monitoring.

Because the origins of disfluency correspond to different stages in speech production, one may expect disfluencies in native speech to follow a non-arbitrary distribution. Indeed, studies on speech production report that hesitations tend to occur before dispreferred or more complex content, such as open-class words (Maclay & Osgood, 1959), unpredictable lexical items (Beattie & Butterworth, 1979), low-frequency color names (Levelt, 1983), or names of low-codability images (Hartsuiker & Notebaert, 2010). The empirical studies introduced above (Arnold et al., 2003, 2007; Barr & Seyfeddinipur, 2010) show that listeners are aware of these regularities in disfluency production: when presented with disfluent speech, listeners anticipated reference to a more cognitively demanding concept. More specifically, listeners attributed disfluencies to speech production difficulties with (i) recognizing unknown objects (e.g., 'I think the speaker is disfluent because she has trouble recognizing the target object'; Arnold et al., 2007; Watanabe et al., 2008) or with (ii) pragmatic status (discourse-new referents in Arnold et al., 2003; Barr & Seyfeddinipur, 2010). These types of attribution involve macroplanning and microplanning, respectively (Levelt, 1989), at the first stage of speech production, namely conceptualization. At this point in the speech production process, the speaker is planning what to say, making use of both the knowledge of the external world and of the discourse model in which the conversation is located (Levelt, 1989; Segalowitz, 2010).

This raises the question how flexible listeners are in attributing the presence of disfluency to other stages in speech production. The current study tests whether listeners may also attribute disfluencies to speech production difficulty further down in the speech production process (Levelt, 1989), namely difficulty in formulation:

> RQ 3A: Do listeners anticipate low-frequency referents upon encountering a disfluency?

Segalowitz (2010) argues that disfluencies may arise as a consequence of the

speaker encountering difficulty in accessing lemma's during the creation of the surface structure (i.e., lexical retrieval). The present three experiments target attribution of disfluency to the speaker having trouble in lexical retrieval by studying the reference resolution of high-frequency (e.g., a hand) vs. low-frequency (e.g., a sewing machine) lexical items. Frequency of occurrence is known to affect lexical retrieval (Almeida, Knobel, Finkbeiner, & Caramazza, 2007; Caramazza, 1997; Jescheniak & Levelt, 1994; Levelt et al., 1999), and, therefore, has been identified as a factor affecting the distribution of disfluencies (Hartsuiker & Notebaert, 2010; Kircher, Brammer, Levelt, Bartels, & McGuire, 2004; Levelt, 1983; Schnadt & Corley, 2006). We hypothesize that, when we present listeners with two known objects, but one having a high-frequency and the other having a low-frequency name, we may find a disfluency bias towards low-frequency objects. Finding a disfluency bias for low-frequency words would extend our knowledge of how disfluencies affect prediction: listeners may attribute disfluencies not only to speaker difficulty with pragmatic status or recognition of unknown objects (conceptualization), but also to difficulty with lexical retrieval of known concepts (formulation). This would be evidence of the competence and efficiency of the predictive mechanisms available to the listener.

Following up on the flexibility of the mechanisms involved in prediction, we know that comprehenders are capable of rapidly modulating the inferences about a speaker's cognitive state based on knowledge about the speaker. In fact, listeners take the speaker's perspective and knowledge into account in reference resolution (Barr & Seyfeddinipur, 2010). The second experiment from Arnold et al. (2007) demonstrated that this latter observation applies to the situation when a listener is convinced he/she is listening to a speaker who suffers from object agnosia. As yet it is unknown whether the disfluency bias is modulated in a much more common situation, namely when listeners are confronted with disfluencies in L2 speech as produced by non-native speakers.

In production, non-natives produce more disfluencies than native speakers do, causing non-native speakers to be perceived as less fluent than native speakers (Cucchiarini et al., 2000, and Chapter 3 of this dissertation). Non-native speech is all the more vulnerable to disfluency due to, for instance, incomplete mastery of the L2 or a lack of automaticity in L2 speech production (Segalowitz & Hulstijn, 2005). This leads to a higher incidence of disfluencies in non-native speech, but it also causes a different distribution of non-native disfluencies (Davies, 2003; Kahng, 2013; Skehan, 2009; Skehan & Foster, 2007; Tavakoli, 2011). While native speakers may produce disfluencies before low-frequency referents due to higher cognitive demands, non-native speakers may experience high cognitive load in naming high-frequency objects as well (e.g., due to poor L2 vocabulary knowledge). As a consequence, the distribution of disfluencies in non-native speech is, from the listener's point of view, more

irregular than the disfluency distribution in native speech. Arguing from this assumption, it follows that non-native disfluencies are, to the listener, worse predictors of the word to follow (as compared to native disfluencies).

We formulate a second research question, addressing the difference between native and non-native disfluencies:

> RQ 3B: Do native and non-native disfluencies elicit anticipation of low-frequency referents to the same extent?

If listeners are aware of the different distribution of disfluencies in non-native speech (as compared to native speech), then we hypothesize that non-native disfluencies will not guide prediction in the same way as native disfluencies. More specifically, hearing a non-native disfluency will not cause listeners to anticipate a low-frequency referent. Thus the disfluency bias may be attenuated when people listen to non-native speech.

Research has shown that exposure to non-native speech can indeed cause the listener to adapt his/her perceptual system. For instance, Clarke and Garrett (2004) found that native English listeners could adapt very rapidly to familiar Spanish-accented speech and to unfamiliar Chinese-accented speech as measured by a decrease in reaction times to visual probe words. Adaptation was shown to take place within one minute of exposure or within as few as two to four utterances. Adaptation to a non-native accent is not only rapid, it is also highly flexible. Bradlow and Bent (2008) reported that, if native listeners are exposed to multiple talkers of the same Chinese accent in L2 English, they could achieve talker-independent adaptation to Chinese-accented English. Hanulíková et al. (2012) investigated the neural correlates of semantic and syntactic violations in native and non-native speech. Semantic violations were observed to result in an N400 effect irrespective of the speaker. This observation suggests that semantic violations in L1 and L2 speech lead to a conflict with the listener's expectations based on (typical) experience: neither native nor non-native speakers are likely to produce sentences with semantic violations. In contrast, grammatical gender violations were observed to result in a P600 effect only when they were produced by a native speaker. When the same violations were produced by a non-native speaker with a foreign accent, no P600 effect was observed. Not only could listeners effectively use a foreign accent as a cue for non-nativeness, moreover, this cue led listeners to adjust their probability model about the grammatical well-formedness of foreign-accented speech. The authors argue that prior experience with non-native speakers producing syntactic errors lies at the core of this cognitive modulation.

The current study investigates the processing of disfluencies in native and non-native speech by means of three eye-tracking experiments. We adopted the experimental procedures of Arnold et al. (2007): studying the disfluency bias in reference resolution by means of the Visual World Paradigm (Huettig et al.,

2011; Tanenhaus et al., 1995). Experiment 1 targets the disfluency bias in native speech. Since this experiment failed in finding evidence for a disfluency bias towards low-frequency referents, an adapted version of this experiment was designed (Experiment 2). Experiment 3 was closely modeled on Experiment 2, but this third experiment makes use of non-native speech materials, thus allowing for a comparison between the processing of native (Experiment 2) and non-native disfluencies (Experiment 3).

The present study not only investigates the disfluency bias as an anticipation effect, but it also targets possible long term effects of disfluency on the processing of the referring expression itself. It has been reported that disfluency may have long term effects on the retention of words in memory. Surprise memory tests following ERP experiments (Collard et al., 2008; Corley et al., 2007; MacGregor et al., 2009, 2010) have revealed that disfluency may have a beneficial effect on the recall accuracy of target words following disfluency. Participants in these studies were presented with a surprise memory test in which participants discriminated between words previously heard in an ERP experiment (old) and words that had not occurred in the ERP experiment (new). Results showed that participants were better at recalling old words when this old word had been preceded by a disfluency. These memory data demonstrate that disfluencies may not only affect online prediction mechanisms, but they may also have long term effects on listeners' information retention. In the present study we study the memory effects of disfluencies in surprise memory tests following eye-tracking experiments. If disfluencies in our eye-tracking experiments have long term effects on the retention of following target words, we expect that listening to disfluencies in native speech (Experiments 1-2) leads to higher recall accuracy of target words. The surprise memory test following Experiment 3 may reveal whether this assumption also holds for non-native disfluencies.

4.2 Experiment 1

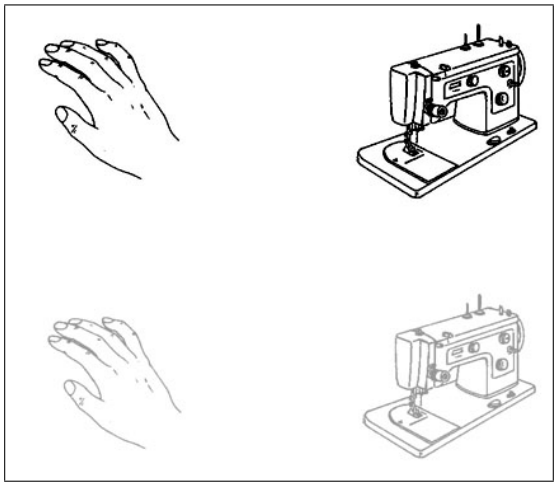
4.2.1 Method of Experiment 1

Participants A sample of 41 participants, recruited from the UiL OTS participant pool, were paid for participation. All participated with implicit informed consent in accordance with local and national guidelines. All were native Dutch speakers and reported to have normal hearing and normal or corrected-to-normal eye-sight ($M_{age} = 21.0$, $SD_{age} = 4.2$, 9m/32f). Data from 3 participants were lost due to technical problems. Data from 6 other participants were excluded from further analyses because their responses on a post-experimental questionnaire indicated suspicion about the experiment (see below). The mean

age of the remaining 32 participants was 21.0 years ($SD_{age} = 4.6$; 7m/25f).

Design and Materials The design of the current experiments resembles that of Arnold et al. (2007). In the Visual World Paradigm as used by Arnold et al. (2007), participants viewed visual arrays on a screen consisting of four pictures: a known object in color A, the same known object in color B, an unknown object in color A and the same unknown object in color B. For an example visual stimulus, see Figure 4.1). The spoken instruction contained a color adjective preceding the target word (e.g., 'Click on thee uh red [target]'), disambiguiting between target and competitor in color A and distractors in color B.

Figure 4.1: Experiment 1: example of high-frequency (hand) and low-frequency (sewing machine) visual stimuli. The top two objects were shown in green, the bottom two in red.



Pictures together with accompanying timed picture naming norms were drawn from the picture set from Severens, Lommel, Ratinckx, and Hartsuiker (2005); see also Appendix C. The black lines of the pictures were replaced by red, green or blue lines using Adobe Photoshop CS5.1. A set of low-frequency (LF; $N = 30$) and a set of high-frequency (HF; $N = 30$) pictures was selected on the basis of log frequencies (as drawn from Severens et al., 2005): mean (SD) log frequency LF=0.38 (0.28); HF=2.07 (0.29); see also Appendix C. An auto-crop procedure was performed on each picture and subsequently its dimensions

were scaled to have a maximum length (or height) of 200 pixels. All pictures selected the Dutch common article *de* (as opposed to the neuter article *het*) and had high name agreement: mean (SD) name agreement LF=96.7 (3.64); HF=97.3 (3.26); see also Appendix C. LF pictures were paired with HF pictures to form a visual array of four pictures for one trial. There was no phonological overlap between the members of these pairs. Together with these experimental pictures, an equal number of LF and HF filler pictures was selected following the same criteria as for the experimental pictures. The only difference between filler and experimental pictures was that half of the filler target objects selected the neuter article *het*. Using a Latin Square design, four pseudo-randomised presentation lists were created. These lists consisted of half LF and half HF targets in both fluent and disfluent instructions (counter-balanced) while disallowing target words to appear in more than one condition.

The audio materials consisted of instructions to click on one of the four objects. These instructions were either fluent or disfluent. A corpus study, based on the Corpus of Spoken Dutch (CGN; Oostdijk, 2000), was conducted to decide on the position of the disfluency in our disfluent sentences. The study targeted the position of the Dutch filled pauses *uh* and *uhm*. It was found that the most common position of Dutch filled pauses was the position preceding the article *de* ($N = 4111$; as compared to the position following the article: $N = 754$). Therefore, the disfluency in our disfluent condition always preceded the article (cf. Arnold et al., 2007, where the disfluency followed the article). For the speech materials of Experiment 1, a female native Dutch speaker (age=21) was recorded. Recordings were made in a sound-attenuated booth using a Sennheiser ME-64 microphone. The speaker was instructed to produce half of the target words (50% HF, 50% LF) in the fluent template (i.e., *Klik op de [color] [target]*, 'Click on the [color] [target]'), and the other half of the target words using a disfluent template, produced 'as naturally as possible' (i.e., *Klik op uh de [color] [target]*, 'Click on uh the [color] [target]'). From all fluent and disfluent sentences that were recorded, six sentence templates (2 fluency conditions x 3 colors) were excised that sounded most natural. These templates extended from the onset of *Klik* to the onset of the color adjective (boundaries set at positive-going zero-crossings, using Praat; Boersma & Weenink, 2012). Additionally, the target words with accompanying color adjectives were excised from the same materials. These target fragments started at the onset of the color adjective at a positive-going zero-crossing. These target fragments were spliced onto a fluent and disfluent sentence template. Thus, target fragments were identical across fluent and disfluent conditions. Since the color adjective was cross-spliced together with the target object, no disfluent characteristics were present in the color or target word.

As a consequence of the described cross-splicing procedure, the differences between fluent and disfluent stimuli were located in the sentence templates

(i.e., fluent *Klik op de*, 'Click on the'; and disfluent *Klik op uh de*, 'Click on uh the'). The instructions were recorded to sound natural. Therefore, apart from the presence of the filled pause *uh*, the contrast between disfluent and fluent stimuli also involved several prosodic characteristics, such as segment duration and pitch (cf. Arnold et al., 2007).

Filler trials were recorded in their entirety; no cross-splicing was applied to these sentences. Instead of counter-balancing the two fluency conditions across the LF and HF filler targets, each LF filler target was recorded in the disfluent condition and each HF filler target was recorded in fluent condition. The reason for this design was that we aimed at a fluent:disfluent ratio across the two frequency conditions which resembled the ratio in spontaneous speech (with disfluencies occurring more often before low-frequency words; Hartsuiker & Notebaert, 2010; Kircher et al., 2004; Levelt, 1983; Schnadt & Corley, 2006). Using our design, the fluent:disfluent ratio was 1:3 for low-frequency targets and 3:1 for high-frequency targets. There was no disfluent template for the disfluent filler trials: they contained all sorts of disfluencies (fillers in different positions, lengthening, corrections, repetitions, etc.).

Apparatus and Procedure Prior to the actual eye-tracking experiment, participants were told a cover story about the purpose of the eye-tracking experiment and about the origins of the speech they would be listening to (following Arnold et al., 2007). Participants were told that recordings had been made of 20 speakers, including both native and non-native speakers of Dutch. Participants in Experiment 1 were told they would be listening to speech from a native speaker. The alleged purpose of the eye-tracking experiment was to test the extent to which instructions from all sorts of speakers could be followed up correctly by listeners. Speakers had purportedly been presented with pictures just like the ones the participant was about to see, but the speaker had seen an arrow appear in the middle of the screen indicating one of the pictures. The speakers' task was then to name that particular picture using a standard instruction template, namely *Klik op de [color] [object]*, 'Click on the [color] [object]'. The presence of the cover story was motivated by the need to justify the presence of disfluencies in the speech. Moreover, it meant that listeners might plausibly attribute the disfluency to difficulty in word retrieval.

Furthermore, participants were familiarized, prior to the eye-tracking experiment, with all the pictures in the experiment ($N = 120$ plus 16 pictures to be used in practice trials) using the ZEP experiment software (Veenker, 2012). Each picture was shown together with its accompanying name (e.g., a picture of a tooth together with the label "tooth"). Participants were instructed to remember the label of each picture. The purpose of this familiarization phase was two-fold: (i) it would help listeners recognize the pictures during the eye-

tracking experiment, and (ii) it would prime the 'correct' name for each picture (e.g., 'tooth', not 'molar'). To ensure participants' attention, participants were presented with test trials after every eighth trial. A test trial involved the depiction of a randomly selected picture from the eight previous pictures. Participants had to type in the correct name for the test picture. When a participant failed to recall the correct label, the test picture was repeated at the end of the familiarization phase. In addition to the 136 pictures to be utilized in the eye-tracking experiment, another set of 30 pictures was added to this familiarization phase which would, in fact, not occur in the eye-tracking experiment. This set was added for use in the surprise memory test (distinguishing between words which had or had not been named during the eye-tracking experiment). Participants were unaware of any difference between the 136 eye-tracking pictures and this extra set of 30 pictures.

In the eye-tracking experiment, eye movements were recorded with a desktop-mounted SR Research EyeLink 1000 eyetracker, controled by ZEP software (Veenker, 2012), which samples the right eye at 500Hz. The system has an eye position tracking range of 32 degrees horizontally and 25 degrees vertically, with a gaze position accuracy of 0.5 degrees. Visual materials were presented on a 19-inch computer screen (within a sound-attenuated eye-tracking booth) at a viewing distance of approximately 60 centimeters. Participants used a standard computer mouse. Speech was heard through speakers at a comfortable listening volume. Before the experiment started, participants were informed about the procedure and the experimenter made sure the participant was comfortably seated. Each experiment started with a thirteen-point calibration procedure followed by a validation procedure. After calibration, participants performed eight practice trials and were given a chance to ask questions. The practice trials contained LF and HF pictures. Two trials contained disfluent speech. A drift correction event occurred before every trial (a red dot appearing in the center of the screen). When the participants had fixated the dot, the two visual stimuli were presented. The onset of the visual stimuli preceded the onset of the audio instructions by 1500 ms. The position of LF and HF picture on the screen was randomized.

Following the eye-tracking experiment, participants were presented with a post-experimental questionnaire. The questionnaire briefly repeated the cover story and, following Barr (2008b), asked participants to rate their level of agreement with four statements on a scale from 1-9 (1 = strong disagreement; 9 = strong agreement). First the naturalness of the speech used in the experiment was assessed. If a participant's response to this first question was lower than 5, it was taken as evidence of suspicion towards the stimuli ($N = 6$, see above). Data from these participants were excluded from further analyses. In any of our three experiments, inclusion of these data did not result in different interpretations of results. The second question elicited accentedness ratings of

the native (Experiment 1-2) and non-native speech (Experiment 3). Thus the 'nativeness' of both speakers, as evaluated by the listeners, could be assessed and compared across experiments. The third question assessed the impression listeners had of the fluency of the speaker. The final question assessed the experience participants had with listening to non-native speakers of Dutch (most relevant for Experiment 3).

Finally, an experimental session finished with a surprise memory test. The purpose of this memory test was to investigate whether target words presented in disfluent contexts had been better remembered than target words presented in fluent contexts. Participants were instructed that they were about to see a set of printed words. Some of these words had and some had not appeared in the eye-tracking experiment. Participants pressed one of two buttons 'as soon as possible while maintaining accuracy' corresponding to whether or not they had heard a particular word in the previous eye-tracking phase. All experimental target words (of which half had been heard in fluent contexts $n = 15$; and half in disfluent contexts $n = 15$) were presented to the participant together with a set of 30 words (15 LF, 15 HF) which had not been part of the previous eye-tracking experiment. In order to avoid a bias towards pictures that had been part of the previous familiarization phase, this set of 30 words had also been added to the familiarization phase (see above). This set was matched to the experimental target words in log frequency of occurrence (as drawn from Severens et al., 2005): mean (SD) log frequency experimental set = 1.23 (0.88); filler set = 1.16 (0.56); $t(58) < 1$. Words were orthographically presented in isolation on the computer screen for 750 ms in a pseudo-random presentation order (with a reversed order counterbalancing any possible order effect). Participants were allowed 2750 ms after word presentation to respond. If no response had been given, the trial was coded as 'incorrect'.

4.2.2 Results of Experiment 1

The reported results follow the order of the experimental sessions: first the eye-tracking data are introduced, followed by the mouse click data, the data from the surprise memory tests, and finally the post-experimental questionnaire.

Eye fixations Prior to the analyses, blinks and saccades were excluded from the data. Eye fixations from trials with a false mouse response were excluded from analyses ($< 1\%$). The pixel dimensions of the object pictures were the regions of interest: only fixations on the pictures themselves were coded as a look toward that particular picture. Eye-tracking data typically contain many missing values. Multilevel analyses are robust against missing data (Quené & Van den Bergh, 2004). Mixed effects logistic regression models (Generalized Linear Mixed Models; GLMMs) as implemented in the `lme4` library (Bates et

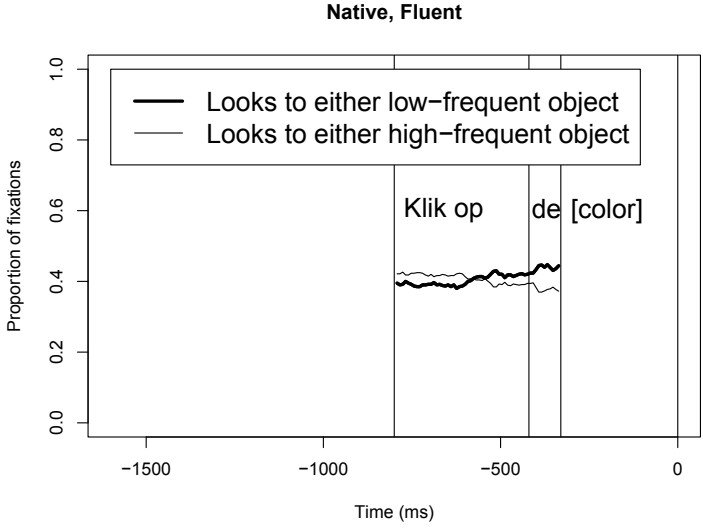
al., 2012) in R (R Development Core Team, 2012) evaluated participants' eye fixations.

Because the present study aimed at finding an *anticipatory* effect triggered by disfluency, the time window of interest should, in any case, precede target onset. Recall that, as a consequence of the described cross-splicing procedure, the differences between fluent and disfluent stimuli were located in the sentence templates. As a consequence, the contrast between disfluent and fluent stimuli involved, next to the presence of the filled pause *uh*, several prosodic characteristics, such as segment duration and pitch (cf. Arnold et al., 2007). Therefore, the left boundary of the time window was set at sentence onset. Finally, because the color adjectives had been recorded in combination with the targets, the color adjectives may have contained some phonetic characteristics of the accompanying target through co-articulation. Therefore, the right boundary of the time window was set at the onset of the color adjective (i.e., at the cross-splicing point). Thus, the time window of interest was defined as starting from sentence onset and ending at the onset of the color adjective (i.e., all fixations while hearing *Klik op de* and *Klik op uh de*). The analyses of the data in this time window tested whether listeners anticipate reference to low-frequency objects in response to disfluency.

Because no phonetic information about the target was available to the listener in the time window of interest, we did not analyse participants' looks to target, as is common in many data analyses of the visual world paradigm. Instead, we analyzed participants' looks to either of the two low-frequency objects. If disfluencies guide prediction, we would expect to find an increase in looks to these two low-frequency objects prior to target onset in the disfluent condition, and not in the fluent condition. Thus, in our GLMMs the dependent variable was the binomial variable LookToLowFrequency (with looks towards either of the two low-frequency objects coded as 'hits', and looks toward high-frequency objects and looks outside the defined regions of interest coded as 'misses'), with participants and items as crossed random effects. Since the time-course of fluent and disfluent trials differed, separate analyses were run per fluency condition, resulting in two separate statistical models. In both models we included a fixed effect of LinearTime, testing for a linear time component (linear increase or linear decrease over time). This factor was centered at *uh* onset in the disfluent model and at 100 ms after sentence onset in the fluent model. All values were divided by 200 in order to facilitate estimation. Furthermore, the factor QuadraticTime ($LinearTime^2$) tested for a quadratic time component (i.e., first an increase followed by a decrease, or first a decrease followed by an increase). Figure 4.2 illustrates the observed looks to the high-frequency and low-frequency objects. The two models are represented in Table 4.1, separately for the fluent and the disfluent model.

Table 4.1 shows that in the fluent condition there were no significant pre-

Figure 4.2: Experiment 1: Proportion of fixations, broken down by fluency. Time in ms is calculated from target onset; note the different time scale of the two panels. The thick lines represent looks to the two low-frequency objects and the thin lines looks to high-frequency objects. Vertical lines represent the (median) onsets of words in the sentence.



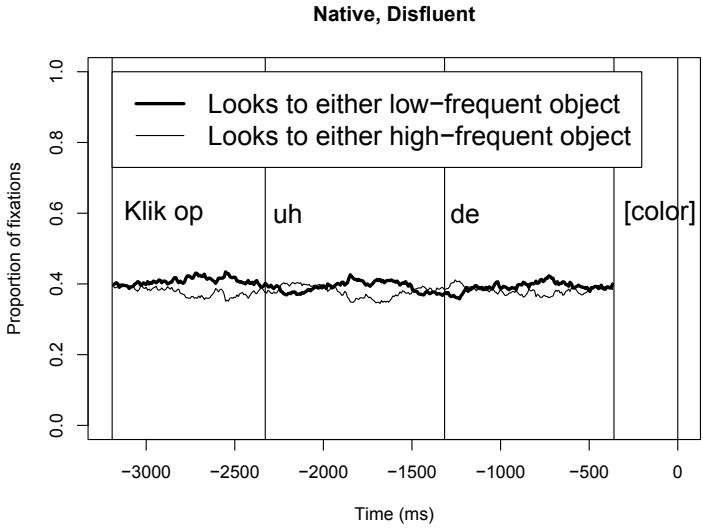


Table 4.1: Experiment 1: Estimated parameters of two mixed effects logistic regression models (standard errors in parentheses; time from sentence onset to the onset of the color adjective) on the looks to low-frequency objects.

FLUENT CONDITION	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(00)}$	-0.697 (0.209)	-3.34	$p < 0.001$ ***
LinearTime, $\gamma_{A(00)}$	0.143 (0.077)	1.87	$p = 0.061$
QuadraticTime, $\gamma_{B(00)}$	0.009 (0.055)	0.17	$p = 0.865$
random effects			
Participant intercept, $\sigma^2_{u_{(j0)}}$	0.935		
Item intercept, $\sigma^2_{v_{(0k)}}$	0.785		
DISFLUENT CONDITION	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(00)}$	-0.565 (0.170)	-3.32	$p < 0.001$ ***
LinearTime, $\gamma_{A(00)}$	-0.011 (0.004)	-3.12	$p = 0.002$ **
QuadraticTime, $\gamma_{B(00)}$	0.001 (0.001)	1.34	$p = 0.180$
random effects			
Participant intercept, $\sigma^2_{u_{(j0)}}$	0.852		
Item intercept, $\sigma^2_{v_{(0k)}}$	0.132		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

dictors. The disfluent model shows a small effect of LinearTime: there was a slight decrease in looks to the two low-frequency pictures across time. These results run counter to our expectation that native disfluencies would elicit a preference for low-frequency referents.

Mouse clicks Participants were very accurate in their mouse clicks (99.6% correct) such that tests for effects of fluency (fluent vs. disfluent) or frequency (low-frequency targets vs. high-frequency targets) on accuracy were not viable. The mouse reaction times (RTs) are given in Table 4.2 (calculated from target onset and for correct trials only). We performed Linear Mixed Effects Regression analyses (LMM; Baayen et al., 2008; Quené & Van den Bergh, 2004, 2008) as implemented in the `lme4` library (Bates et al., 2012) in R (R Development Core Team, 2012) to analyze the mouse click RTs (log-transformed). The random effects in this model consisted of the factor Participant, testing for individual differences between participants; Item, testing for differences between items; and Order, testing for individual differences in order effects, varying within participants. More complex random effects did not significantly improve the model. The fixed part of the model consisted of the factor IsDisfluent, testing for differences between fluent and disfluent trials; and IsLowFrequency, testing for differences between trials with a high-frequency vs. a low-frequency

Table 4.2: Experiment 1: Mean reaction times of mouse clicks (in ms, calculated from target onset and for correct trials only; standard deviation in brackets).

	Native speech	
	Fluent	Disfluent
High-frequency target	1006 (314)	984 (285)
Low-frequency target	1040 (298)	1017 (317)

target object. Interactions between two fixed effects were also added as predictors. Finally, a fixed effect of Order tested for any order effects. The number of degrees of freedom required for statistical significance testing of t values was given by $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 32$ participants) and m is the total number of explanatory variables in the model ($m = 8$) resulting in 23 degrees of freedom. This statistical model revealed that none of the predictors reached significance.

Surprise memory test The recall accuracy and reaction times of participants' responses in the surprise memory test are represented in Table 4.3. Reaction times were calculated from word presentation onwards and for correct trials only. First we analysed the recall accuracy. We tested a mixed effects logistic regression model (Generalized Linear Mixed Model; GLMM) with random effects consisting of the factor Participant, testing for individual differences between participants, and Item, testing for differences between items. More complex random effects did not significantly improve the model. The fixed part of the model consisted of the previously introduced factors IsDisfluent, IsLowFrequency, and a fixed Order effect. Interactions between IsDisfluent and IsLowFrequency were also added as predictors. A main effect of IsLowFrequency was found to significantly affect the recall accuracy ($p = 0.037$): participants in both experiments were significantly more accurate recalling low-frequency objects as compared to high-frequency objects. There was neither a main effect of IsDisfluent, nor any interaction of this factor with IsLowFrequency. Similar statistical testing on the reaction times from the surprise memory test revealed no significant effects.

Post-experimental questionnaire Participants had rated the naturalness, the accentedness, and the fluency of the speech stimuli on a scale from 1-9 (with higher ratings indicating more natural, more accented, more fluent speech). The average naturalness of the speech was rated 6.37 ($SD = 1.59$). The average accentedness of the stimuli was rated 1.00 ($SD = 0$). The fluency

Table 4.3: Experiment 1: Mean recall accuracy (in percentages) and mean reaction times (in ms from word presentation onwards, correct trials only) of participants' responses (standard deviation in brackets).

	Native speech	
	Fluent	Disfluent
recall accuracy		
High-frequency target	55 (50)	59 (49)
Low-frequency target	64 (48)	69 (46)
reaction times		
High-frequency target	1002 (346)	1030 (325)
Low-frequency target	978 (312)	1003 (307)

of the speech was rated 5.39 ($SD = 1.67$) and the extent to which participants regularly interacted with non-native speakers of Dutch in their daily lives was rated 3.88 ($SD = 2.34$).

4.2.3 Discussion of Experiment 1

The eye-tracking data from Experiment 1 only revealed a very small linear decrease in looks to the two low-frequency objects, found in the time window preceding the onset of the color adjective. Closer inspection of the eye-tracking data that preceded the time window of interest (i.e., during the 1500 ms that the visual stimuli were displayed without any audio instructions) revealed a consistent 'novelty' preference for the low-frequency objects at the onset of visual stimulus presentation. The slight decrease in looks to the two low-frequency objects may indicate a decrease of the novelty of the low-frequency objects as time progressed. In any case, these data do not support our expectation that native disfluencies would elicit anticipation of low-frequency referents. Furthermore, no disfluency effects were found in the mouse click data, nor in the surprise memory test. Several factors may be thought to be responsible for these null effects. First of all, we included a familiarization phase in our experimental design to prime the correct labels for the pictures used in the eye-tracking experiment. However, this familiarization phase may have reduced the contrast between high-frequency and low-frequency pictures in the eye-tracking experiment because both types of pictures had been recently viewed by the participants. Secondly, the time between the disfluency *uh* and the point of disambiguation (i.e., target onset) is relatively long in the experimental design of Experiment 1. Finding a disfluency bias for low-frequency referents in the current experimental design would involve listeners having to maintain their

expectation of a low-frequency referent for a lengthy period of time. This may be unlikely considering the relative weak effect of disfluencies on reference resolution (cf. Arnold et al., 2007). In fact, a re-analysis of the looks to the low-frequency pictures in a smaller time window, namely from *uh* onset to *de* offset, did reveal a significant effect of QuadraticTime which was only present in the disfluent condition (i.e., an increase followed by a decrease in looks to the low-frequency picture, only in the disfluent condition). Taken together, these observations argue for designing a new experiment with a smaller time span between the disfluency and target onset.

Therefore, a second experiment was designed. In this second experiment, the familiarization phase was removed from the experimental design. The high name agreement of the pictures (mean name agreement LF=96.7; HF=97.3) was thought to be sufficient for participants to activate the correct label for each of the pictures. Furthermore, the time between the disfluency *uh* and target onset was reduced by removing the color adjective from the stimulus sentences: instead of hearing *Klik op uh de [color] [target]*, 'Click on uh the [color] [target]' in the disfluent condition, the stimulus sentence in Experiment 2 was reduced to *Klik op uh de [target]*, 'Click on uh the [target]'. Because the colors were removed from the audio instructions, the number of visual referents on the screen was reduced to two: one black line-drawing of a high-frequency object and one black line-drawing of a low-frequency object.

The third experiment was identical to Experiment 2 except that Experiment 3 tested the perception of non-native disfluencies. Therefore, in Experiment 3, participants listened to a non-native speaker of Dutch producing fluent and disfluent instructions with a strong foreign accent. Comparing the results from Experiment 2-3 may reveal differential effects of native and non-native disfluencies on the predictive mechanisms involved in speech perception. First the method of Experiment 2 is outlined, below, followed by the similar method of Experiment 3. Subsequently, the statistical analyses involving the data from both Experiment 2 and 3 are described.

4.3 Experiment 2

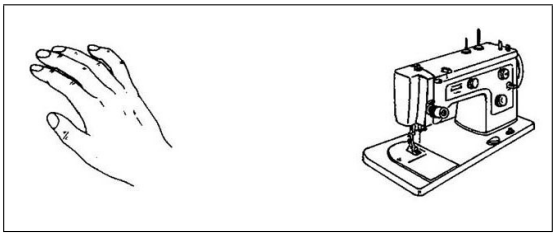
4.3.1 Method of Experiment 2

Participants A sample of 44 participants, recruited from the UiL OTS participant pool, were paid for participation. All participated with implicit informed consent in accordance with local and national guidelines. All were native Dutch speakers and reported to have normal hearing and normal or corrected-to-normal eye-sight ($M_{age} = 23.7$, $SD_{age} = 8.1$, 13m/31f). Data from 3 participants were lost due to technical problems. Data from 6 other participants were

excluded from further analyses because their responses on a post-experimental questionnaire indicated suspicion about the experiment (see below). The mean age of the remaining 35 participants was 23.8 years ($SD_{age} = 8.4$; 11m/24f).

Design and Materials The design of Experiment 2 resembled that of Experiment 1. However, where Experiment 1 did include a familiarization phase, no such familiarization phase was present in Experiment 2. Moreover, Experiment 2 used visual arrays consisting of only two objects (one low-frequency, one high-frequency; see Figure 4.3). The pictures from Experiment 1 were re-used for Experiment 2.

Figure 4.3: Experiment 2-3: Example of a picture pair, consisting of one high-frequency (hand) and one low-frequency object (sewing machine), used in both experiments.



The audio materials of Experiment 2 consisted of instructions to click on one of the two objects. These instructions were either fluent or disfluent. For the speech materials of Experiment 2, a female native Dutch speaker (age=30) was recorded. Recordings were made in a sound-attenuated booth using a Sennheiser ME-64 microphone. The speaker was instructed to produce half of the target words (50% HF, 50% LF) in the fluent template (i.e., *Klik op de [target]*, 'Click on the [target]'), and the other half of the target words using a disfluent template, produced 'as naturally as possible' (i.e., *Klik op uh de [target]*, 'Click on uh the [target]'). From all fluent and disfluent sentences that were recorded, six sentence templates (three recordings of each fluency condition) were excised that sounded most natural. These templates extended from the onset of *Klik* to the onset of the article *de* (boundaries set at positive-going zero-crossings, using Praat; Boersma & Weenink, 2012). The target words were excised from the same materials. These target fragments started at the onset of the article *de* at a positive-going zero-crossing and were spliced onto a fluent and disfluent sentence template. Thus, target words were identical across fluent and disfluent conditions.

As a consequence of the described cross-splicing procedure, the differences between fluent and disfluent stimuli were located in the sentence templates (i.e., fluent *Klik op*, 'Click on'; and disfluent *Klik op uh*, 'Click on uh'). The instructions were recorded to sound natural. Therefore, apart from the presence of the filled pause *uh*, the contrast between disfluent and fluent stimuli also involved several prosodic characteristics (cf. Arnold et al., 2007). For instance, the words *Klik op*, 'Click on', in the disfluent condition were longer and had a higher pitch as compared to the fluent condition (see Table 4.4 for prosodic properties of the native and non-native sentence templates).

Table 4.4: Experiment 2-3: Duration (in ms) and pitch (in Hz) for the three fluent and three disfluent sentence templates in the native and non-native speech.

	Klik	op	uh
	Native speech		
fluent			
Duration	194, 199, 214	147, 166, 180	
Maximum pitch	217, 220, 227	214, 222, 237	
disfluent			
Duration	213, 218, 262	245, 264, 283	871, 889, 933
Maximum pitch	261, 262, 282	260, 269, 270	244, 246, 263
	Non-native speech		
fluent			
Duration	214, 221, 221	191, 195, 198	
Maximum pitch	225, 228, 237	227, 230, 255	
disfluent			
Duration	221, 240, 261	234, 254, 263	891, 897, 950
Maximum pitch	273, 278, 287	282, 287, 304	259, 273, 280

Filler trials were recorded in their entirety; no cross-splicing was applied to these sentences. Instead of counter-balancing the two fluency conditions across the LF and HF filler targets, each LF filler target was recorded in the disfluent condition and each HF filler target was recorded in fluent condition (identical to Experiment 1). The reason for this design was that we aimed at a fluent:disfluent ratio across the two frequency conditions which resembled the ratio in spontaneous speech (with disfluencies occurring more often before low-frequency words; Hartsuiker & Notebaert, 2010; Kircher et al., 2004; Levelt, 1983; Schnadt & Corley, 2006). Using our design, the fluent:disfluent ratio was 1:3 for low-frequency targets and 3:1 for high-frequency targets. There was no disfluent template for the disfluent filler trials: they contained all sorts of disfluencies (*uhm*'s in different positions, lengthening, corrections, etc.).

Apparatus and Procedure The procedure of Experiment 2 was identical to that of Experiment 1, except that there was no familiarization phase.

4.4 Experiment 3

4.4.1 Method of Experiment 3

Experiment 3 was identical to Experiment 2 except that non-native speech was used.

Participants A new sample of 42 participants, recruited from the UiL OTS participant pool, were paid for participation. All participated with implicit informed consent in accordance with local and national guidelines. All were native Dutch speakers and reported to have normal hearing and normal or corrected-to-normal eye-sight ($M_{age} = 22.7$, $SD_{age} = 3.2$, 5m/37f). Data from 6 participants were excluded because their responses on a post-experimental questionnaire indicated suspicion about the experiment (having provided naturalness ratings below 5 in the post-experimental questionnaire). The mean age of the remaining 36 participants was 22.6 years ($SD_{age} = 3.3$), 5m/31f.

Design and Materials The visual stimuli were identical to those used in Experiment 2. For the speech materials of Experiment 3, a non-native speaker of Dutch was recorded (female, L1 Romanian, age=25, LoR=3.5 years). She reported having rudimentary knowledge of Dutch (self-reported CEFR level A1/A2) and limited experience using Dutch in daily life. Recordings were made in a sound-attenuated booth using a Sennheiser ME-64 microphone. In order to have a minimal contrast between the native and non-native recordings, we adopted the recording procedures of Hanulíková et al. (2012): the non-native speaker first listened to a native utterance after which she imitated the native speech, sentence by sentence. This resulted in non-native speech recordings that were identical to the native recordings except for a noticeable foreign accent (see Table 4.4 for prosodic properties of the native and non-native speech stimuli). This procedure was adopted for both the experimental and the filler trials. The remaining procedure was identical to Experiment 1.

Apparatus and Procedure The cover story, the instructions, the post-experimental questionnaire and the surprise memory test were identical to Experiment 2, except that participants in Experiment 3 were instructed that they were going to listen to a *non-native* speaker of Dutch.

4.5 Results from Experiment 2-3

Data from both experiments were combined in all analyses. The reported results follow the order of the experimental sessions: first the eye-tracking data are introduced, followed by the mouse click data, the data from the surprise memory tests, and finally the post-experimental questionnaire.

4.5.1 Eye fixations

Prior to the analyses, blinks and saccades were excluded from the data. Eye fixations from trials with a false mouse response were excluded from analyses ($< 1\%$). The pixel dimensions of the object pictures were the regions of interest: only fixations on the pictures themselves were coded as a look toward that particular picture. The eye-tracking data were analyzed using Generalized Linear Mixed Models (GLMMs), similar to the analyses of Experiment 1.

The eye fixation data were evaluated in two time windows: one *pre-target time window* preceding article onset and one *post-target time window* following article onset. Note that the time windows refer to the time in which (i) there was no target information available (pre-target time window preceding the splicing point) and the time in which (ii) the target description was presented (post-target time window following the splicing point). Thus, the analyses of the data in the pre-target time window tested whether listeners anticipate, *prior to target onset*, reference to low-frequency objects following disfluency. Analyses of the post-target time window were carried out to test for any spillover effects onto the eye fixations following target onset.

Pre-target time window The time window of interest was defined as starting from sentence onset and ending before article onset (i.e., all fixations while hearing *Klik op* and *Klik op uh*). In the pre-target time window no phonetic information about the target was available to the listener. Therefore, we did not analyse participants' looks to target, as is common in many data analyses of the visual world paradigm. Instead, we analyzed participants' looks to the low-frequency object. If disfluencies guide prediction, we would expect to find an increase in looks to low-frequency objects prior to target onset in the disfluent condition, and not in the fluent condition. Thus, in our GLMMs the dependent variable was the binomial variable LookToLowFrequency (with looks towards low-frequency objects coded as 'hits', and looks toward high-frequency objects and looks outside the defined regions of interest coded as a 'misses'), with participants, items, and sentence templates as three crossed random effects. Since the time-course of fluent and disfluent trials differed, separate analyses were run per fluency condition.

In both models we included (1) a fixed effect of IsNonNative, to test for differences between native and non-native speech; (2) a fixed effect of LinearTime, testing for a linear time component (linear increase or linear decrease over time). This factor was centered at *uh* onset in the disfluent model and at 100 ms after sentence onset in the fluent model. All values were divided by 200 in order to facilitate estimation. Furthermore, (3) the factor QuadraticTime ($LinearTime^2$) tested for a quadratic time component (i.e., first an increase followed by a decrease, or first a decrease followed by an increase). Also, the interactions between the factor IsNonNative and the two time components were included in both models. We also tested for a cubic time component, which significantly improved the fit of the model of the disfluent data. However, the addition of a cubic time component did not lead to a qualitatively different interpretation of results. For the sake of intelligibility, we only present models without a cubic time component here. Figure 4.4 illustrates the combined linear, quadratic, and interaction effects of IsNonNative and time on the estimated proportion of looks to low-frequency objects. The two models are represented in Table 4.5, separately for the fluent and the disfluent model.

Inspection of the first model (data from the fluent condition in the upper panel) reveals that there were no effects of IsNonNative or any time component: there was no preference for either of the two pictures. Inspection of the second model (data from the disfluent condition in the lower panel) reveals that several predictors affected the likelihood of a look toward a low-frequency picture when listeners were presented with a disfluent sentence. The predictor LinearTime demonstrates that there was an increase in looks toward low-frequency pictures across time. The predictor QuadraticTime reveals that there was a negative quadratic time component in the disfluent data. This indicates an increase in looks toward low-frequency pictures followed by a decrease. The interactions of the time components with the factor IsNonNative reveals that the significant effects of the time components only applied to the data from Experiment 2: only when listeners were presented with native disfluent speech did we find a preference for looking toward low-frequency pictures. These results confirm our expectation that native disfluencies elicit anticipation of low-frequency referents, but non-native disfluencies do not.

The graphs in Figure 4.4 illustrate the preference for low-frequency objects in the native disfluent condition. The rise in looks to low-frequency objects in the native disfluent condition starts before the median onset of the disfluency *uh*. Two factors may account for this early rise. Firstly, there was some variance in the onset of the disfluency across the three disfluent sentence templates, but this variance was not very large (maximal negative deviance from the median: -130 ms). Secondly, the early preference may be due to the disfluent character of the disfluent sentence template as a whole, including the prosodic characteristics of the content preceding the filled pause *uh* (see Table 4.4).

Figure 4.4: Experiment 2-3: Proportion of fixations, broken down by fluency and nativeness, in the pre-target time window. Time in ms is calculated from target onset. Vertical lines represent the (median) onsets of words in the sentence.

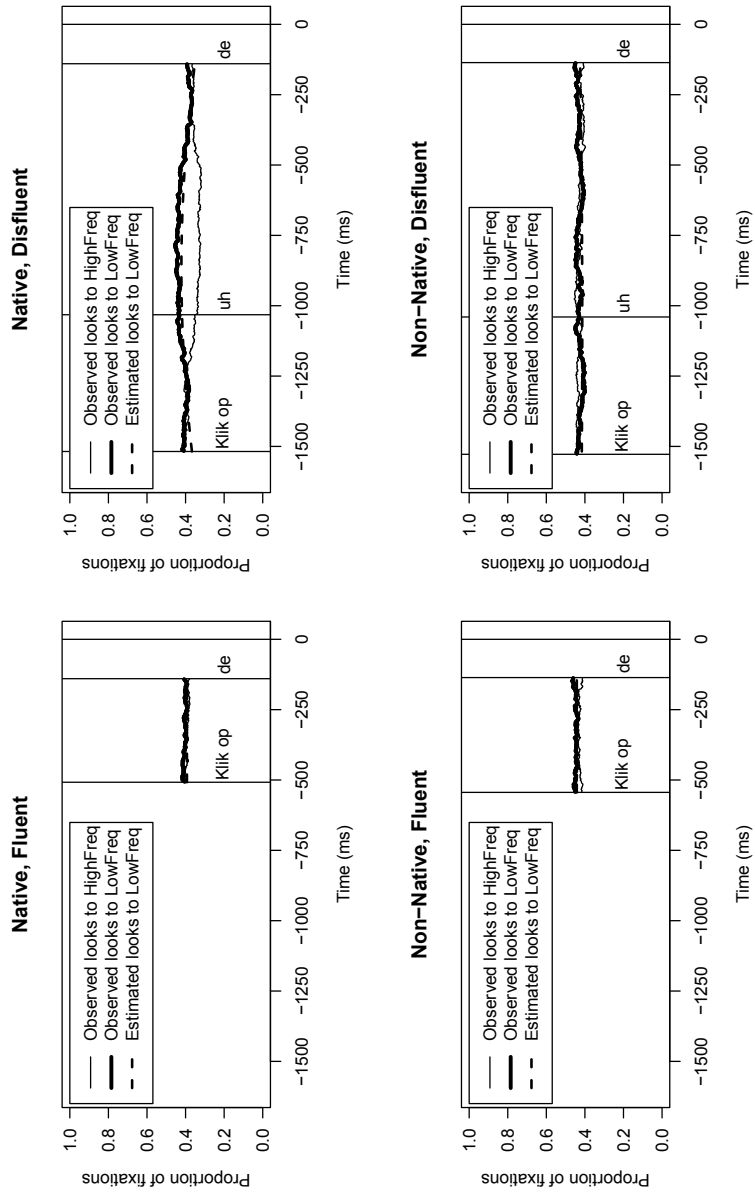


Table 4.5: Experiment 2-3: Estimated parameters of two mixed effects logistic regression models (standard errors in parentheses; pre-target time window from sentence onset to article onset) on the looks to low-frequency objects.

MODEL OF FLUENT CONDITION	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(000)}$	-0.632 (0.201)	-3.15	$p = 0.002$ **
LinearTime, $\gamma_{B(000)}$	-0.041 (0.029)	-1.41	$p = 0.160$
QuadraticTime, $\gamma_{C(000)}$	0.018 (0.031)	0.58	$p = 0.565$
IsNonNative, $\gamma_{A(000)}$	0.399 (0.252)	1.58	$p = 0.113$
IsNonNative x LinearTime, $\gamma_{D(000)}$	-0.039 (0.039)	-1.00	$p = 0.319$
IsNonNative x QuadraticTime, $\gamma_{E(000)}$	0.049 (0.038)	1.29	$p = 0.197$
random effects			
Participant intercept, $\sigma^2_{u_{(j00)}}$	1.050		
Item intercept, $\sigma^2_{v_{(0k0)}}$	0.237		
Sentence template intercept, $\sigma^2_{w_{(00l)}}$	0.006		
MODEL OF DISFLUENT CONDITION	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(000)}$	-0.330 (0.142)	-2.32	$p = 0.020$ *
LinearTime, $\gamma_{B(000)}$	0.035 (0.003)	12.51	$p < 0.001$ ***
QuadraticTime, $\gamma_{C(000)}$	-0.022 (0.001)	-21.22	$p < 0.001$ ***
IsNonNative, $\gamma_{A(000)}$	-0.019 (0.187)	-0.10	$p = 0.920$
IsNonNative x LinearTime, $\gamma_{D(000)}$	-0.034 (0.004)	-8.72	$p < 0.001$ ***
IsNonNative x QuadraticTime, $\gamma_{E(000)}$	0.024 (0.001)	16.86	$p < 0.001$ ***
random effects			
Participant intercept, $\sigma^2_{u_{(j00)}}$	0.316		
Item intercept, $\sigma^2_{v_{(0k0)}}$	0.052		
Sentence template intercept, $\sigma^2_{w_{(00l)}}$	0.025		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Post-target time window Analyses of the post-target time window were carried out to test for any spillover effects onto the eye fixations following target onset. Visual inspection of the data in the post-target time window revealed that participants correctly looked at target within 500 ms of target onset. Thus the time window of interest was defined from article onset to 500 ms after target onset. In this time window participants heard phonological information about the target object. Therefore, in contrast to the pre-target time window, here we analyzed participants' looks to target. If disfluencies guide prediction, we would expect listeners to identify the low-frequency target object faster in the disfluent condition relative to the fluent condition. Conversely, we may also find high-frequency targets to be recognized slower in the disfluent condition.

In our GLMMs the dependent variable was the binomial variable LookToTarget (with looks towards target objects coded as 'hits', and looks toward competitor objects and looks outside the defined regions of interest coded as 'misses'), with participants, items, and sentence templates as three crossed random effects. In the post-target time window, the time-course was identical across conditions because the spoken realizations of article and target were identical due to cross-splicing. Therefore, one large analysis on the data from both experiments was run including the aforementioned predictors IsNonNative, LinearTime (centered around target onset), and QuadraticTime ($LinearTime^2$). Additionally, the predictor IsLowFrequency, testing for differences between trials with a high-frequency vs. a low-frequency target object, and the predictor IsDisfluent, testing for differences between the fluent and disfluent condition, were included in the fixed part of the model. Finally, the interactions between the factor IsNonNative, IsDisfluent, IsLowFrequency and the two time components were included in the model. Again, a cubic time component significantly improved the fit of the model but for simplicity we only present a model without a cubic time component. If the anticipation of low-frequency referents following disfluency, found for the data from Experiment 2, spills over to the post-target time window, we would expect to find a significant four-way interaction between IsNonNative, IsLowFrequency, IsDisfluent, and one of the time components. Figure 4.5 illustrates the estimated linear, quadratic, and interaction effects across the fluency and frequency conditions, separately for the native and non-native data. The statistical model is represented in Table 4.6.

Visual inspection of Figure 4.5 suggests that, for the native data from Experiment 2 in the top panel, participants' looks to high-frequency target words following a disfluency were distinct from the other conditions (cf. the thick dashed line from 0-200 ms in the top panel of Figure 4.5). Listeners looked less at high-frequency targets (i.e., more at the low-frequency competitor) when they had heard a disfluency precede the target description. In the lower panel of Figure 4.5, the non-native data from Experiment 3, there does not seem to be any difference between thick (disfluent trials) and thin lines (fluent trials).

Figure 4.5: Experiment 2-3: Estimated proportion of fixations on target, broken down by fluency, target frequency and nativeness, in the post-target time window. Time in ms is calculated from target onset. Vertical lines represent the (median) onsets of words in the sentence.

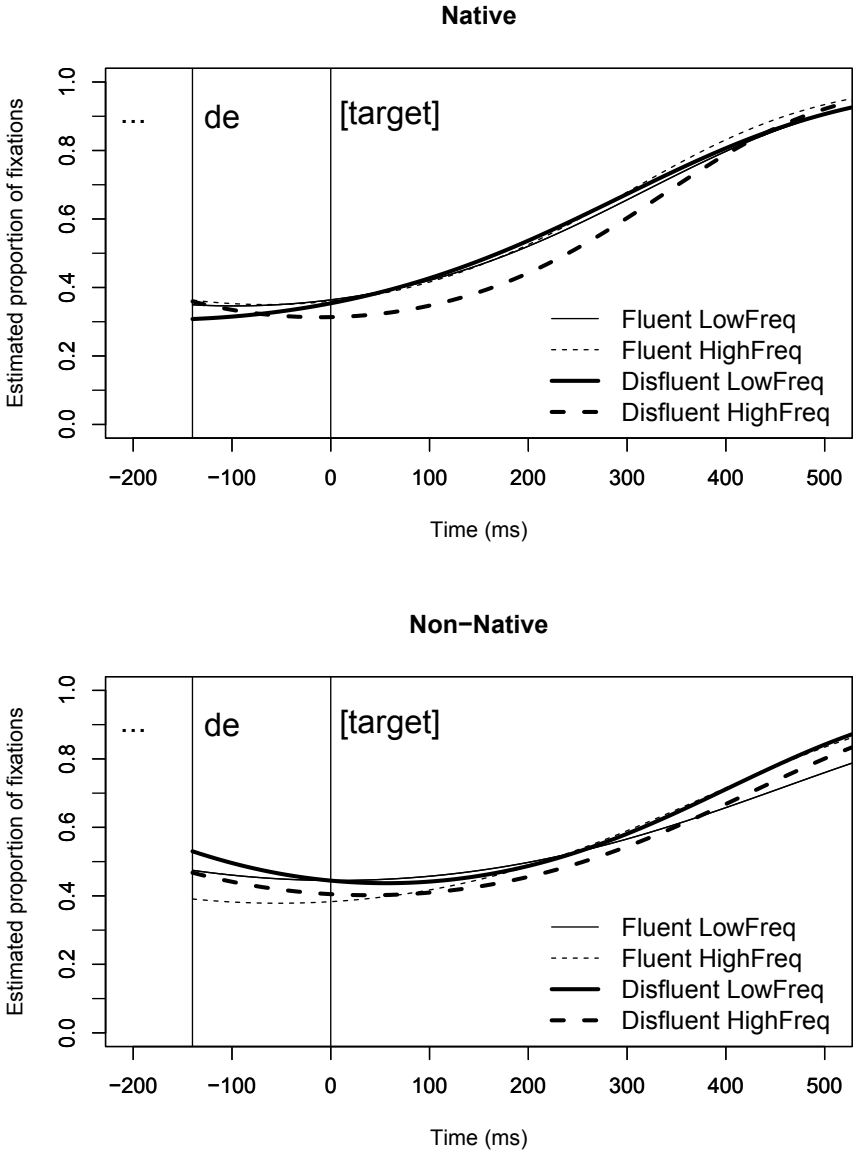


Table 4.6: Experiment 2-3: Estimated parameters of mixed effects logistic regression modelling (standard errors in parentheses; post-target time window from article onset to 500 ms after target onset) on the looks to target.

	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(000)}$	-0.575 (0.121)	-4.77	$p < 0.001$ ***
LinearTime, $\gamma_{A(000)}$	0.267 (0.013)	20.18	$p < 0.001$ ***
QuadraticTime, $\gamma_{B(000)}$	0.410 (0.008)	53.16	$p < 0.001$ ***
IsLowFrequency, $\gamma_{C(000)}$	0.017 (0.011)	1.50	$p = 0.133$
IsLowFrequency x LinearTime, $\gamma_{D(000)}$	0.050 (0.019)	2.66	$p = 0.008$ **
IsLowFrequency x QuadraticTime, $\gamma_{E(000)}$	-0.086 (0.011)	-8.02	$p < 0.001$ ***
IsDisfluent, $\gamma_{F(000)}$	-0.280 (0.134)	-2.08	$p = 0.037$ *
IsDisfluent x LinearTime, $\gamma_{G(000)}$	-0.070 (0.026)	-2.62	$p = 0.009$ **
IsDisfluent x QuadraticTime, $\gamma_{H(000)}$	0.016 (0.015)	1.06	$p = 0.290$
IsDisfluent x IsLowFrequency, $\gamma_{I(000)}$	0.165 (0.022)	7.42	$p < 0.001$ ***
IsDisfluent x IsLowFrequency x LinearTime, $\gamma_{J(000)}$	0.378 (0.037)	10.12	$p < 0.001$ ***
IsDisfluent x IsLowFrequency x QuadraticTime, $\gamma_{K(000)}$	-0.145 (0.021)	-6.84	$p < 0.001$ ***
IsNonNative, $\gamma_{L(000)}$	0.099 (0.156)	0.63	$p = 0.526$
IsNonNative x LinearTime, $\gamma_{M(000)}$	-0.120 (0.018)	-6.55	$p < 0.001$ ***
IsNonNative x QuadraticTime, $\gamma_{N(000)}$	-0.133 (0.010)	-12.84	$p < 0.001$ ***
IsNonNative x IsLowFrequency, $\gamma_{O(000)}$	0.239 (0.015)	15.47	$p < 0.001$ ***
IsNonNative x IsLowFrequency x LinearTime, $\gamma_{P(000)}$	-0.209 (0.026)	-8.01	$p < 0.001$ ***
IsNonNative x IsLowFrequency x QuadraticTime, $\gamma_{Q(000)}$	0.034 (0.014)	2.37	$p = 0.018$ *
IsNonNative x IsDisfluent, $\gamma_{R(000)}$	0.300 (0.190)	1.58	$p = 0.114$
IsNonNative x IsDisfluent x LinearTime, $\gamma_{S(000)}$	-0.064 (0.037)	-1.76	$p = 0.079$
IsNonNative x IsDisfluent x QuadraticTime, $\gamma_{T(000)}$	-0.028 (0.021)	-1.41	$p = 0.181$
IsNonNative x IsDisfluent x IsLowFrequency, $\gamma_{U(000)}$	-0.189 (0.031)	-6.07	$p < 0.001$ ***
IsNonNative x IsDisfluent x IsLowFrequency x LinearTime, $\gamma_{V(000)}$	-0.450 (0.052)	-8.61	$p < 0.001$ ***
IsNonNative x IsDisfluent x IsLowFrequency x QuadraticTime, $\gamma_{W(000)}$	0.322 (0.029)	11.18	$p < 0.001$ ***
random effects			
Participant intercept, $\sigma^2_{u_{(J00)}}$	0.275		
Item intercept, $\sigma^2_{v_{(0K0)}}$	0.066		
Sentence template intercept, $\sigma^2_{w_{(00I)}}$	0.027		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Rather, listeners look more at low-frequency target objects (solid lines) than at high-frequency targets (dashed lines) in both fluent and disfluent conditions.

The model described in Table 4.6 statistically tests the data of Figure 4.5. Because the model is quite complex, we have split the fixed effects of the model into two parts: the upper part is comprised of predictors that are related to Experiment 2 (native speaker), the lower part involves predictors that are related to Experiment 3 (main effect of IsNonNative, and interactions). We will first inspect the upper part of the model.

The first eleven predictors ($\gamma_{A(000)}$ - $\gamma_{K(000)}$) apply to the native data from Experiment 2. The model took fluent trials with high-frequency targets as its the intercept. Thus, the first two predictors ($\gamma_{A(000)}$ and $\gamma_{B(000)}$) show significant effects of the linear and quadratic time component in fluent trials with a high-frequency target: there was an overall increase in looks to target ($\gamma_{A(000)}$) and this increase accumulated quadratically ($\gamma_{B(000)}$); cf. the thin dashed line in the top panel of Figure 4.5.

Predictors $\gamma_{C(000)}$ - $\gamma_{E(000)}$ compare, within the fluent condition, trials with a high-frequency target to trials with a low-frequency target. We observe a slightly stronger linear increase and a slightly weaker quadratic time component in this condition (cf. the thin solid line in the top panel of Figure 4.5).

The following three predictors ($\gamma_{F(000)}$ - $\gamma_{H(000)}$) apply to disfluent trials with a high-frequency target (cf. the thick dashed line in the top panel of Figure 4.5). In this condition, disfluency negatively affected target recognition: there were considerably fewer looks to high-frequency targets at target onset ($\gamma_{F(000)}$) and a somewhat weaker increase in looks to target ($\gamma_{G(000)}$).

Finally, the interactions between IsDisfluent, IsLowFrequency, and the time components ($\gamma_{I(000)}$ and $\gamma_{K(000)}$) show that disfluency positively affected the recognition of low-frequency targets (cf. the thick solid line in the top panel of Figure 4.5): listeners looked more at low-frequency targets at target onset ($\gamma_{I(000)}$) and the linear increase over time was stronger ($\gamma_{J(000)}$). A negative effect of the quadratic time component ($\gamma_{K(000)}$) showed that in disfluent trials with a low-frequency target the increase in looks to target was more linear than in the other conditions. That is, where participants in the other conditions were somewhat slower in looking to target as indicated by the quadratic nature of the increase in looks, participants were faster in looking to target in disfluent trials with a low-frequency target.

Judging from the upper part of the fixed effects, the main observation that was established was that participants in Experiment 2 (listening to a native speaker) looked less at high-frequency targets (i.e., more at the low-frequency competitor) when they had heard a disfluency precede the target description. The lower part of the fixed effects ($\gamma_{L(000)}$ - $\gamma_{W(000)}$) investigates the looking behaviour of participants in Experiment 3 (listening to a non-native speaker).

The first three predictors ($\gamma_{L(000)}$ and $\gamma_{N(000)}$) apply to the intercept con-

dition: fluent trials with a high-frequency target (cf. the thin dashed line in the lower panel of Figure 4.5). The linear and quadratic time components in this condition in the native data ($\gamma_{A(000)}$ and $\gamma_{B(000)}$) are observed to be somewhat weaker in the non-native data.

Predictors $\gamma_{O(000)}$ - $\gamma_{Q(000)}$ compare the intercept condition to fluent trials with a low-frequency target (cf. the thin solid line in the lower panel of Figure 4.5). At target onset listeners looked more at target when this target was low-frequency ($\gamma_{O(000)}$). An even more negative effect of the linear time component ($\gamma_{P(000)}$) and a positive effect of the quadratic time component ($\gamma_{Q(000)}$) showed that in fluent non-native trials the increase in looks to target was more quadratic, where the increase was more linear for high-frequency targets.

The following three predictors ($\gamma_{R(000)}$ - $\gamma_{T(000)}$) apply to disfluent non-native trials with a high-frequency target (cf. the thick dashed line in the lower panel of Figure 4.5). No statistically significant effects of IsDisfluent were found for the non-native data. Finally, the interactions between IsNonNative, IsDisfluent, IsLowFrequency, and the time components ($\gamma_{U(000)}$ and $\gamma_{W(000)}$) show that disfluency negatively affected the recognition of low-frequency targets (cf. the thick solid line in the lower panel of Figure 4.5): listeners looked less at low-frequency targets at target onset ($\gamma_{U(000)}$) and the linear increase over time was considerably weaker ($\gamma_{V(000)}$). The positive effect of the quadratic time component ($\gamma_{W(000)}$) indicated that in disfluent trials with a low-frequency target the increase in looks to target was more quadratic. Summing up, the lower part of the fixed effects, testing the looking behaviour of participants in Experiment 3 (listening to a native speaker), did not reveal any interaction between disfluency and participants' preference for either of the two objects (in contrast to the looking behaviour of participants in Experiment 2).

Revisiting Figure 4.5, we observe that the deviation of the 'Disfluent + High-Frequency' condition in the native data is located in the first 400 ms following target onset. It is estimated that planning and executing a saccade takes approximately 100-200 ms (see Altmann, 2011, for review). Taking this estimate into account, participants initially anticipated reference to a low-frequency object (from 0-200 ms). However, when the first phonetic details of the unexpected high-frequency target became available to the listeners (roughly from 200 ms onwards), listeners moved their eyes away from the anticipated low-frequency object (i.e., fixating the unexpected high-frequency object at approximately 400 ms). These results demonstrate spillover effects of the anticipation in the pre-target time window, found for the data from Experiment 2, to the eye fixations in the post-target time window.

Note that in Figure 4.4 and Figure 4.5 there seems to be a higher baseline in the bottom panels picturing the non-native data from Experiment 3. This observation is based on visual inspection alone, since we did not find a significant effect of IsNonNative in any of our statistical models.

4.5.2 Mouse clicks

Across the two experiments, participants were very accurate in their mouse clicks (Experiment 2: 99.7%; Experiment 3: 100%) such that tests for effects of fluency (fluent vs. disfluent) or frequency (low-frequency targets vs. high-frequency targets) on accuracy were not viable. The mouse reaction times (RTs) are given in Table 4.7 (calculated from target onset and for correct trials only). We performed Linear Mixed Effects Regression analyses (LMM; Baayen et al., 2008; Quené & Van den Bergh, 2004, 2008) as implemented in the `lme4` library (Bates et al., 2012) in R (R Development Core Team, 2012) to analyze the mouse click RTs (log-transformed) from both experiments. The random effects in this model consisted of the factor Participant, testing for individual differences between participants; Item, testing for differences between items; and Order, testing for individual differences in order effects, varying within participants. More complex random effects did not significantly improve the model.

The fixed part of the model consisted of the factor IsNonNative, testing for differences between native and non-native speech; the factor IsDisfluent, testing for differences between fluent and disfluent trials; and IsLowFrequency, testing for differences between trials with a high-frequency vs. a low-frequency target object. Interactions between these three fixed effects were also added as predictors. The number of degrees of freedom required for statistical significance testing of t values was given by $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 30$ experimental items) and m is the total number of explanatory variables in the model ($m = 11$) resulting in 18 degrees of freedom. Three predictors were found to significantly affect the RTs: (1) a main effect of IsNonNative ($p = 0.017$) revealed that participants listening to a non-native speaker responded slower than participants listening to a native speaker; (2) a main effect of IsLowFrequency ($p < 0.001$) revealed that participants were slower responding to LF targets relative to HF targets; and (3) an interaction between IsDisfluent and IsLowFrequency ($p = 0.036$) counteracted the negative effect of IsLowFrequency: when low-frequency targets were presented in disfluent context, participants were slightly faster in their response than when the low-frequency target was presented in fluent context.

4.5.3 Surprise memory test

The recall accuracy and reaction times of participants' responses in the surprise memory test are represented in Table 4.8. Reaction times were calculated from word presentation onwards and for correct trials only. First we analysed the recall accuracy across the two experiments. We tested a mixed effects logistic regression model (Generalized Linear Mixed Model; GLMM) with random effects

Table 4.7: Experiment 2-3: Mean reaction times of mouse clicks (in ms, calculated from target onset and for correct trials only) in both experiments (standard deviation in brackets).

	Native speech		Non-native speech	
	Fluent	Disfluent	Fluent	Disfluent
High-frequency target	774 (244)	792 (214)	870 (277)	892 (236)
Low-frequency target	849 (267)	832 (260)	954 (271)	962 (301)

consisting of the factor Participant, testing for individual differences between participants, and Item, testing for differences between items. More complex random effects did not significantly improve the model. The fixed part of the model consisted of the previously introduced factors IsNonNative, IsDisfluent, and IsLowFrequency. Interactions between these three fixed effects were also added as predictors. A main effect of IsLowFrequency was found to significantly affect the recall accuracy ($p < 0.001$): participants in both experiments were significantly more accurate recalling low-frequency objects as compared to high-frequency objects. There was neither a main effect of IsDisfluent, nor any interaction of this factor with IsNonNative or IsLowFrequency. Similar statistical testing on the reaction times from the surprise memory tests (in both experiments) revealed no significant effects.

Table 4.8: Experiment 2-3: Mean recall accuracy (in percentages) and mean reaction times (in ms from word presentation onwards, correct trials only) of participants' responses in both experiments (standard deviation in brackets).

	Native speech		Non-native speech	
	Fluent	Disfluent	Fluent	Disfluent
recall accuracy				
High-frequency target	54 (50)	51 (50)	60 (49)	59 (49)
Low-frequency target	67 (47)	71 (45)	75 (44)	74 (44)
reaction times				
High-frequency target	854 (271)	868 (274)	892 (295)	825 (280)
Low-frequency target	839 (309)	842 (243)	832 (266)	860 (267)

4.5.4 Post-experimental questionnaire

Participants in both experiments had rated the naturalness, the accentedness, and the fluency of the speech stimuli on a scale from 1-9 (with higher ratings indicating more natural, more accented, more fluent speech). The average naturalness of the speech was rated 7.05, $SD = 1.73$ (native) and 6.12,

$SD = 1.77$ (non-native), $t(83) = 2.44, p = 0.017$. The average accentedness of the stimuli was rated 1.44, $SD = 1.33$ (native) and 6.10, $SD = 1.90$ (non-native), $t(83) = -13.11, p < 0.001$. The fluency of the speech from both experiments was rated 5.88, $SD = 2.11$ (native) and 5.36, $SD = 1.82$ (non-native), $t(83) = 1.23, p = 0.221$. Finally, participants also rated the extent to which they regularly interacted with non-native speakers of Dutch in their daily lives: 4.00, $SD = 1.99$ (native) and 3.83, $SD = 2.13$ (non-native), $t(83) < 1$.

4.6 General discussion

Our first eye-tracking experiment failed to establish a native disfluency bias for low-frequency referents. However, the adjustments in Experiment 2 revealed that listeners may attribute disfluency to speaker trouble with lexical retrieval. We attribute this difference between the results of Experiment 1 and Experiment 2 to the absence of a familiarization phase in Experiment 2, and shorter stimulus sentences in Experiment 2.

When participants in Experiment 2 were presented with native disfluent speech, they fixated low-frequency objects more than high-frequency objects. This effect was observed in the pre-target time window, indicating anticipation of low-frequency referents upon encountering a disfluency. This anticipation effect persisted into the post-target time window, where it surfaced as a dispreference for high-frequency targets in the native disfluent condition. The effects observed in the eye-tracking data were confirmed by the mouse click reaction times: participants were faster to click on a low-frequency target when this target was preceded by a disfluency. Taken together, our results suggest that listeners are sensitive to the increased likelihood of speakers to be disfluent while referring to low-frequency objects (Hartsuiker & Notebaert, 2010; Kircher et al., 2004; Levelt, 1983; Schnadt & Corley, 2006). Moreover, this sensitivity guides them to use disfluency as a cue to predict reference to a low-frequency object. This finding extends our understanding of the comprehension system. It has been shown that listeners may use disfluencies to guide prediction of dispreferred or more complex linguistic content. For instance, listeners may predict discourse-new (Arnold et al., 2003; Barr & Seyfeddinipur, 2010) or unknown referents (Arnold et al., 2007) upon hearing a disfluency. In the fluency framework of Segalowitz (2010), this involves attribution of disfluency to conceptualization: comprehenders infer that the speaker is having trouble with planning what to say, integrating both knowledge of the external world and of the current discourse model. Our experiments involved pictures that were all familiar, but differed in the frequency of occurrence of the lexical items. Therefore, listeners could not have attributed disfluency to difficulty in conceptualization, but rather to difficulty in formulation of speech. Our study

demonstrates that listeners use disfluencies to infer that the speaker is encountering difficulty at another stage in speech production, namely lexical retrieval. This finding emphasizes the flexibility of the language architecture, particularly of the predictive mechanisms available to the listener.

Comparing our results (attribution of disfluency to formulation) with those from Arnold et al. (2007) (attribution of disfluency to conceptualization), we find that the magnitude of the disfluency bias varies. In Arnold et al. (2007) the preference for unknown referents was somewhat stronger (maximal difference in proportion of looks between fluent and disfluent condition: approximately 20%) than the disfluency bias reported in our pre-target time window (maximal difference: approximately 10%). This difference may be related to the different dimensions tested: the probability of disfluency preceding reference to completely unknown and unidentifiable objects (as in Arnold et al., 2007) may be higher than the probability of disfluency occurring before reference to known, but low-frequency, objects. This difference in probability may have led listeners to have a stronger preference, upon hearing a disfluency, for unknown referents (Arnold et al., 2007) than for low-frequency referents (this study).

Note that the disfluency bias, observed in the eye-tracking data from Experiment 2, surfaced both in the pre-target time window and in the post-target time window. Similar results were found in the study by Arnold et al. (2003). There, the authors interpreted the disfluency bias in the pre-target time window as anticipation of discourse-new referents. The fact that the disfluency bias persisted in their post-target window was interpreted as disfluency facilitating the identification of the referential expression itself. However, Barr and Seyfeddinipur (2010) state that such interpretations may be misleading because they confound effects that emerge during the post-target time window with anticipation effects that may have emerged earlier and that persist over the time window (Barr, 2008a, 2008b). Therefore, our disfluency bias in the post-target time window may be interpreted as a spillover effect of the disfluency bias observed in the pre-target time window.

In fact, we aimed at finding longer term effects of disfluency by means of our surprise memory tests, but no disfluency effects on the retention of target words were observed. Previous surprise memory tests indicated a beneficial effect of disfluency on the recognition probability of the following target noun (e.g., Corley et al., 2007; MacGregor et al., 2010). The data from the present surprise memory tests did not show a beneficial effect of disfluency, only of target frequency: higher recall accuracy of low-frequency words relative to high-frequency words. The surprise memory tests reported in previous studies, evaluated participants' recall accuracy of stimuli presented in ERP experiments, whereas our memory tests investigated recall of stimuli presented in eye-tracking experiments. Owing to this difference, the lack of a disfluency effect may be attributed to several factors. For instance, the memory tests

reported by Corley and colleagues differed from our tests in the duration of experimental sessions, the total number of trials, and the linguistic content of the speech stimuli. Any of these factors may be responsible for the null result obtained here. Our data only warrant the conclusion that disfluencies, in native speech, affect the prediction of target words, but no support was found for disfluency facilitating the identification or retention of referential expressions themselves.

Experiment 3 allowed for a comparison between the processing of native and non-native disfluencies. When listeners were presented with native speech containing disfluencies (Experiment 2), a disfluency bias for low-frequency referents was observed. In contrast, when listeners were presented with non-native speech (Experiment 3), the disfluency bias for low-frequency referents was absent: no difference was found between the fluent and disfluent non-native speech conditions. Thus we extend the reported attenuation of the disfluency bias when people listen to a speaker with object agnosia (Arnold et al., 2007, Experiment 2) to a much more common situation, namely when people listen to a non-native speaker. Recall that the non-native speaker, in producing the non-native speech materials, had imitated the native speech stimuli (following the method from Hanulíková et al., 2012). As a consequence, the non-native materials closely resembled the native speech materials (see, for instance, Table 4.4). The principal difference between the native and non-native stimuli was the presence of a foreign accent in the non-native speech (average accent rating of 6.1 on a 9-point scale). Therefore, the attenuation of the disfluency bias in Experiment 3 can be attributed to the listeners' perception of a foreign accent. Listeners can effectively use a foreign accent as a cue for non-nativeness and adjust their predictions accordingly (cf. Hanulíková et al., 2012). These adjustments do not necessarily affect behavioral measures of listeners' speech comprehension. Disfluency was found to speed up participants' mouse clicks to low-frequency targets, irrespective of whether participants were listening to native or non-native speech (no interactions between IsNonNative and IsDisfluent was observed).

Observing a difference between the processing of native and non-native disfluencies, raises the question what the source of this difference might be. It seems that listeners' prior experiences with non-native speech modified their expectations about the linguistic content following disfluencies. L2 speech production is cognitively more demanding than producing L1 speech (De Bot, 1992; Segalowitz, 2010). As a consequence, the incidence and the distribution of disfluencies in L2 speech is different from that in L1 (Davies, 2003; Kahng, 2013; Skehan, 2009; Skehan & Foster, 2007; Tavakoli, 2011).

This difference between the native and non-native distribution of disfluencies may be argued to be the result of non-native speakers experiencing high cognitive load where a native speaker would not (i.e., due to the fact that the

non-native speaker is speaking in his L2). In fact, the 'weaker links hypothesis', as proposed by Gollan, Montoya, Cera, and Sandoval (2008), argues that the limited exposure to L2 words makes them, for an L2 speaker, functionally equivalent to L1 low-frequency words. Thus, lexical retrieval of high-frequency lexical items may be just as cognitively demanding for a non-native speaker as lexical retrieval of low-frequency lexical items would be for a native speaker. Therefore, from the native listener's point of view, the distribution of disfluencies in non-native speech is more irregular than the disfluency distribution in native speech.

The results from Experiment 3 indicate that listeners take non-native disfluencies to be worse predictors of the word to follow and, therefore, the effect of non-native disfluencies on prediction is attenuated. This may involve modification of the probability model about speech properties. Brunelliére and Soto-Faraco (2013) propose that L1 listeners have less specified phonological expectations when listening to non-native speech, based on prior experience with the irregular phonology of L2 speakers. Analogous to less specified phonological expectations, L1 listeners may adjust their probability model about the linguistic content following a non-native disfluency in response to prior experience with the irregularities of non-native disfluency production. Note that these adjustments are stereotype-dependent: on the basis of the discernment of a foreign accent, listeners draw inferences about the L2 proficiency of the non-native speaker. Apparently, listeners bring stereotypes to bear for speech comprehension, when perceiving certain voice characteristics (Van Berkum, Van den Brink, Tesink, Kos, & Hagoort, 2008).

This raises the question whether the effect of such stereotypes (e.g., of non-native speakers) on speech comprehension may be modulated. For instance, how would listeners respond to hearing a non-native speaker whom they know to be a very proficient L2 speaker? It remains to be seen whether the attenuation of the disfluency bias when listening to non-native speech is a gradual process that can be affected by the inferred proficiency of the non-native speaker. Furthermore, our results do not necessarily preclude non-native disfluencies from guiding prediction in all situations. This would only hold if listeners take the distribution of non-native disfluency production to be too arbitrary to make any kind of reliable prediction. Our data show that non-native disfluencies do not guide listeners to anticipate reference to low-frequency objects. Further investigation will have to unravel whether listeners make use of non-native disfluencies to anticipate other types of referents, such as discourse-new or unknown objects (i.e., attribution to speaker trouble in conceptualization).

In conclusion, the present study contributes to the notion that comprehenders are adept at making linguistic predictions. Not only do listeners anticipate certain linguistic content on the basis of linguistic representations of the utterance (e.g., semantics, syntax, phonology), but also on the basis of performance

characteristics, that is, disfluency. Moreover, the current data highlight the adaptable nature of the comprehension system in two ways. Firstly, listeners are capable of attributing symptoms of inefficiency in speech production (i.e., disfluencies) to difficulty in conceptualization of unknown referents (Arnold et al., 2007) or to difficulty in formulation (i.e., lexical retrieval) of low-frequency referents. Secondly, when listeners have knowledge about the non-native identity of the speaker, these attributions may be modulated as evidenced by attenuation of predictive strategies. Previous studies indicate that knowledge about the speaker may affect listeners' comprehension in a range of ways. A sentence in a situation of speaker inconsistency (e.g., hearing a male speaker utter the improbable sentence 'I am pregnant') may elicit larger N400 effects than the same sentence in a speaker consistent condition (e.g., spoken by a female speaker; Van Berkum et al., 2008). Hearing a non-native speaker produce syntactic errors elicits a smaller P600 effect than the same error produced by a native speaker (Hanulíková et al., 2012). The current experiments showed that hearing a foreign accent influences the way listeners use performance aspects of the speech signal to guide prediction. Taken together, these studies emphasize the central role of speaker characteristics in comprehension and prediction.

CHAPTER 5

Do L1 and L2 disfluencies heighten listeners' attention?

5.1 Introduction

Although engaging in conversation is a common acticity, producing fluent speech is strikingly difficult. Speakers have to decide on the conceptual message they want to convey, find a formulation of the message, and articulate the appropriate sounds (Levelt, 1989). Moreover, all these cognitive processes are to be executed in a timely fashion since conversation takes place at a remarkable speed. Therefore, it is not surprising that speakers often have to stall for time by means of hesitations, such as silent and filled pauses (e.g., *uh*'s and *uhm*'s).

Hesitations, or disfluencies, have been defined as "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (Fox Tree, 1995), such as silent pauses, filled pauses, corrections, repetitions, etc. It has been estimated that six in every hundred words are affected by disfluency (Bortfeld et al., 2001; Fox Tree, 1995). Segalowitz (2010) proposed, in his fluency framework adapted from Levelt (1989) and De Bot (1992), that the (dis)fluent character of an utterance is defined by the speaker's *cognitive fluency*: the operation efficiency of speech planning, assembly, integration and execution. If the efficiency of the speech production process falters, disfluencies in the utterance are the result.

Empirical work on speech production has shown that the aforementioned definition of disfluencies is, to some extent, incomplete. Disfluencies may not

add propositional content to an utterance, but they do cue information about the linguistic content following disfluency. Disfluencies in spontaneous speech have been found to follow a non-arbitrary distribution. Because disfluency in the speech signal may arise as a result of speaker trouble in speech production, disfluencies tend to occur before open-class words (Maclay & Osgood, 1959), unpredictable lexical items (Beattie & Butterworth, 1979), low-frequency color names (Levelt, 1983), or names of low-codability images (Hartsuiker & Notebaert, 2010). Hesitations, therefore, cue the onset of dispreferred or more complex content.

But do listeners actually make use of disfluencies as cues to more complex information? Several perception studies have targeted the effects that disfluencies have on speech comprehension, converging on the conclusion that listeners are sensitive to the distribution of disfluencies. The perception literature indicates that listeners use the increased likelihood of speakers to be disfluent before more complex information (1) to predict the linguistic content following disfluency, and (2) to raise their attention levels to the following linguistic content.

Evidence for disfluency effects on prediction comes from eye-tracking and ERP research. ERP studies show that listeners integrate unpredictable target words more easily into a disfluent context than a fluent context (Corley et al., 2007; MacGregor et al., 2010), as evidenced by an attenuation of the N400 effect in disfluent sentences. Eye-tracking studies report that, upon encountering the filled pause *uh* in a sentence such as 'Click on thee uh [target]', listeners are more likely to look at pictures of discourse-new objects (Arnold et al., 2003, 2004; Barr & Seyfeddinipur, 2010), unidentifiable objects (Arnold et al., 2007; Watanabe et al., 2008), or low-frequency lexical items (Chapter 4 of this dissertation). This suggests that listeners use disfluency as a cue to predict the relative complexity of the linguistic content to follow.

The link between listeners' experience with the non-arbitrary distribution of disfluencies, on the one hand, and disfluency effects on prediction, on the other hand, was emphasized in Chapter 4 of this dissertation. Here, it was argued that, in contrast to the non-arbitrary distribution of disfluencies in native speech, non-native speakers produce disfluencies in much more irregular patterns. Non-native speech is vulnerable to disfluency due to the fact that non-native speakers experience high cognitive load in (L2) speech production much more frequently (compared to native speakers). This leads non-native speakers to produce more disfluencies than native speakers and it causes a different distribution of non-native disfluencies (Davies, 2003; Kahng, 2013; Skehan, 2009; Skehan & Foster, 2007; Tavakoli, 2011). From the point of view of the listener, the distribution of disfluencies in non-native speech is more irregular than the distribution of disfluencies in native speech.

Moreover, research seems to indicate that listeners are aware of the differ-

ent distribution of non-native disfluencies. The experiments in Chapter 4 of this dissertation report that listeners were found to attenuate the use of non-native disfluencies for prediction. Where participants listening to native speech were observed to have a disfluency bias for low-frequency referents (i.e., upon encountering a disfluency, there were more looks to pictures of low-frequency objects [e.g., a sewing machine] than to pictures of high-frequency objects [e.g., a hand]), no such disfluency bias could be established when participants listened to a non-native speaker with a strong foreign accent. This suggests that listeners are aware of the more irregular patterns of disfluencies in non-native speech, and, therefore, modulate the effect of non-native disfluencies on prediction.

Disfluencies do not only guide prediction; they have also been observed to trigger listeners' attention. Three partially distinct functional components of attention have been identified, namely orienting, detecting targets, and maintaining alert states (Posner & Petersen, 1990). Collard (2009) has argued that disfluencies provide the listener with auditory novelty that triggers an *orienting* response (disengagement, shift, reengagement). He has reported evidence of disfluency affecting listeners' attention by making use of the Change Detection Paradigm (CDP).

In this Change Detection Paradigm, participants listen to speech passages which they try to remember. After listening to the speech, a textual representation of the passage is presented which either matches the spoken passage or contains a one word substitution. Participants have the task to indicate through a button press whether they detect a change in the text or not. In the CDPs reported in Collard (2009), the to-be-substituted words (i.e., target words) in the spoken passages were either presented in a fluent context or a disfluent context, with a filled pause (e.g., *uh*) preceding the target word. Collard (2009) found that listeners were more accurate at detecting a change in a CDP when the target word had been encountered in the context of a hesitation (relative to presenting the target word in a fluent speech passage). As such, the Change Detection Paradigm can be used to show that disfluencies trigger listeners' attention, with consequences for the retention of the following words.

There have been several other studies that have targeted participants' recall of previously presented words. For instance, Corley et al. (2007) and MacGregor et al. (2010) tested participants on their recall of words previously presented in ERP experiments. They found that participants were more accurate in recalling words that had been preceded by a disfluency than words that had been presented in a fluent sentence. Fraundorf and Watson (2011) found that listeners were better at recalling plot points from previously remembered stories, when these stories contained filled pauses (as compared to disfluency-free stories). A beneficial effect of disfluency was observed across plot points, regardless of whether one particular plot point had contained a disfluency or not.

More direct evidence of heightened attention levels being responsible for

the memory effects of disfluencies, comes from an ERP study by Collard et al. (2008). Participants in this study listened to sentences that sometimes contained a sentence-final target word that had been acoustically compressed, thus perceptually deviating from the rest of the sentence. This acoustic deviance induced ERP components associated with attention (mismatch negativity [MMN] and P300). However, when the deviant target word was preceded by a disfluency, the P300 effect was strongly reduced. This suggests that listeners were not required to reorient their attention to deviant words in disfluent cases. Moreover, a surprise memory test established, once again, a beneficial effect of disfluency on the recognition of previously heard words.

It could be argued that the disfluency effects on attention have the same origins as the disfluency effects on prediction. Because disfluency introduces novel, dispreferred or more complex information, listeners may benefit from anticipating more complex linguistic content *and* from raising their attention as a precautionary measure to ensure timely comprehension of the unexpected information. Thus, the regularities in the distribution of disfluencies would be responsible for the disfluency effects on both prediction and attention: due to their non-arbitrary distribution, disfluencies elicit anticipation of more complex information, and trigger listeners' attention. Heightened attention, then, affects the recognition and retention of words following the disfluency. Following up on this assumption, one could expect *non-native* disfluencies to have differential effects on listeners' attention. The distribution of non-native disfluencies has been argued, above, to be more irregular than the native distribution. As such, raised attention levels in response to non-native disfluencies may not prove advantageous to the native listener. Therefore, listeners may modulate the effect of non-native disfluencies on attention.

Alternatively, the effects of disfluencies on attention may be the result of more automatic cognitive processes in response to delay. Corley and Hartsuiker (2011) have proposed a *Temporal Delay Hypothesis* accounting for beneficial effects of disfluencies on auditory word recognition. They argue that it is not necessary to postulate listener sensitivity to the distributional properties of speech following disfluencies. Instead, temporal delay - inherent to disfluency - facilitates listeners' recognition and listeners' retention of words. Support for this hypothesis comes from studies that have compared effects of different types of delays on word recognition (RTs) and word retention (recall accuracy). For instance, filled pauses have been reported to speed up word recognition (i.e., lower RTs for words following filled pauses; Brennan & Schober, 2001; Corley & Hartsuiker, 2011; Fox Tree, 2001), but similar effects have been reported for silent pauses and sine tones (Corley & Hartsuiker, 2011). However, conflicting results were found by Fraundorf and Watson (2011) who showed that filled pauses had a beneficial effect on listeners' recall of story plot points, but coughs (matched in duration to the filled pauses) did not.

These two explanations of disfluency effects on attention lead to different predictions when it comes to non-native disfluency. If attentional effects are automatically triggered due to delay, then both native and non-native delay should result in heightened attention levels. This would suggest that the disfluency effects on attention are more automatic than the disfluency effects on prediction (which may be modulated on the basis of knowledge about the non-native identity of the speaker; Chapter 4 of this dissertation). If, however, the attentional effects are a consequence of the distribution of disfluencies, then non-native disfluencies - with their more irregular distribution - might not affect attention in the same way as native disfluencies do. In fact, we may find an attenuation of attentional effects when it comes to non-native disfluency. In the literature we find support for rapid modulation of the listener's perceptual system on the basis of knowledge about the non-native identity of the speaker (e.g., Hanulíková et al., 2012, Chapter 4 of this dissertation).

The present study consists of two experiments addressing the following research question:

> RQ 4: Do native and non-native disfluencies trigger heightened attention to the same extent?

Our first experiment targets the effect of disfluencies in native speech on listeners' attention. For this, we adopt the Change Detection Paradigm (CDP) from Collard (2009, Experiment 3): participants indicate whether a written transcript matches a previously heard spoken passage or not (i.e., contains a one word substitution). Crucially, the to-be-substituted words (i.e., target words) in the spoken passages are presented either in a fluent context or a disfluent context, with a filled pause (e.g., *uh*) preceding the target word. We hypothesize that we replicate the results from Collard (2009, Experiment 3) for Dutch: listeners are predicted to be more accurate at detecting a change in our CDP when the changed word had been preceded by a filled pause. The beneficial effect of disfluency on participants' accuracy in the CDP is taken to be indicative of increased attention triggered by disfluency.

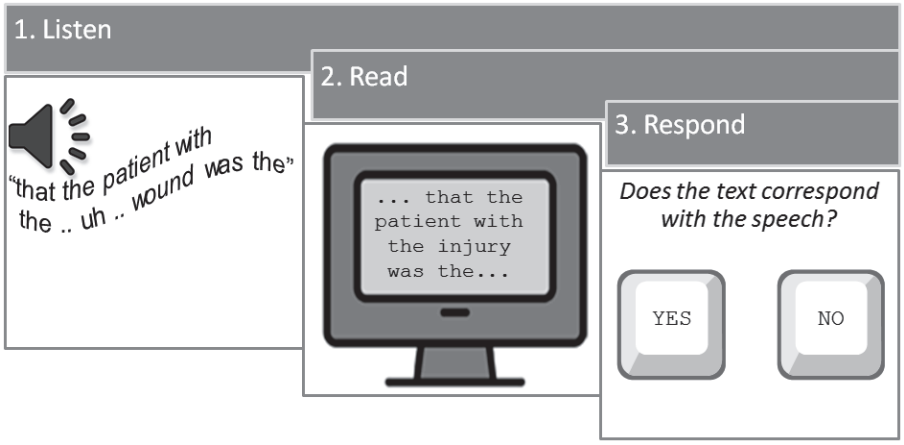
The second experiment investigates whether listeners modulate their attentional mechanisms in response to non-native disfluency. Instead of using native speech materials, participants in Experiment 2 listened to a non-native speaker producing the same fluent and disfluent passages from Experiment 1. If non-native disfluencies trigger listeners' attention to the same extent as native disfluencies, this would provide support for an automatic-processing account of the attentional effects of disfluency. Conversely, if non-native disfluencies do not trigger listeners' attention, then this would suggest that listeners' attentional mechanisms may be modulated by the (more irregular) distribution of non-native disfluencies.

5.2 Method

5.2.1 Experiment 1

The method of Experiments 1 and 2 was adapted from the Change Detection Paradigm (CDP; schematically represented in Figure 5.1) described in Experiment 3 of Collard (2009).

Figure 5.1: Schematical representation of the Change Detection Paradigm. Example of the CloseChange condition.



Participants A sample of 40 participants participated in Experiment 1 with implicit informed consent in accordance with local and national guidelines. All were native Dutch speakers and reported to have normal hearing ($M_{age} = 22.3$, $SD_{age} = 2.3$, 7m/33f).

Design and Materials A sample of 36 experimental passages was adopted from Collard (2009), each passage consisting of three sentences (see Appendix D). An experimental trial involved the presentation of a recording of one passage that was either fluent or disfluent marked by a filled pause (e.g., Table 5.1). The word following the disfluency is refered to as the target word.

Table 5.1: Example passage adapted from Collard (2009).

Dutch example passage: De dokter keek hoe lang hij nog moest werken. Hij zag dat de patiënt met de [uh] *wond* als enige nog in de wachtkamer zat. De vriendelijke maar strikte verpleegster bracht de jongen de spreekkamer binnen.
English translation: The doctor checked to see how much longer he had to work. He saw that the patient with the [uh] *wound* was the only one present in the waiting room. A kind but strict nurse brought the boy into the consulting room.

We used three types of change conditions:

- NoChange condition: text passage identical to spoken passage (e.g., *wound* → *wound*).
- DistantChange condition: text passage contains a substitution involving a semantically unrelated noun (e.g., *wound* → *handkerchief*).
- CloseChange condition: text passage contains a substitution involving a semantically related noun (e.g., *wound* → *injury*).

The three change conditions differed with respect to the written transcript that was presented after the audio passage. This written text was either identical to the speech previously heard (the NoChange condition) or it contained a one word substitution. This substitution involved either a change to a semantically related noun (e.g., *wound* → *injury*; CloseChange) or to a word that was not related to the original target word (e.g., *wound* → *handkerchief*; DistantChange). Target words (e.g., *wound* in Table 5.1) were always located in the second sentence of a passage, in a prepositional phrase that was out of focus. The frequency of occurrence of NoChange (e.g., *wound*), CloseChange (e.g., *injury*), and DistantChange words (e.g., *handkerchief*) was obtained from SUBTLEX-NL, a database of Dutch word frequencies based on 44 million words from film and television subtitles (Keuleers et al., 2010). Words were matched in the log-transformed frequency of occurrence per million words (mean (SD): NoChange 0.915 (0.848); CloseChange 0.945 (0.725); DistantChange 0.786 (0.830); $F[2, 105] < 1$) and in the number of characters in each word (mean (SD): NoChange 7.0 (2.5); CloseChange 7.1 (2.8); DistantChange 7.8 (3.5); $F[2, 105] < 1$).

For the speech materials of Experiment 1, a male native speaker of Dutch (age=25) was recorded. Recordings were made in a sound-attenuated booth using a Sennheiser ME-64 microphone (16-bit, 44000Hz, mono). The speaker was

instructed to speak as clearly as possible and to make the disfluencies sound as natural as possible. Two recordings were made per passage: one fluent passage and one disfluent passage with a filled pause preceding the target word (see Table 5.1). In order to make the distinction between the two fluency conditions as minimal as possible, the stimuli used for the actual experiment were created through speech manipulation (using Praat; Boersma & Weenink, 2012). First, the initial and final sentences were extracted from the recordings. Secondly, a new fluent version of the second sentence was created by excising the filled pause from the disfluent version (at positive-going zero-crossings). If removing the disfluency led to an unnatural result, we instead inserted the disfluency from the disfluent condition into the fluent sentence. This second procedure was required for three passages. Concatenating the first, second (fluent or disfluent) and third sentence resulted in our experimental stimuli (36 audio passages in 2 conditions). In this fashion, we made sure that the two versions of each passage (fluent vs. disfluent) were identical except for the presence of the filled pause in the disfluent version.

To avoid the participants becoming accustomed to the positions of the targets, or disfluencies, or the co-occurrence of the two, 18 filler passages were included in the experiment. Filler passages were recorded in their entirety; no cross-splicing was applied to these sentences. Half of the filler passages were presented in the DistantChange condition, the other half in the NoChange condition (i.e., there was no CloseChange condition in the filler trials). If a change occurred in a filler trial, then this change never occurred in the second sentence. Half of the filler trials contained a disfluency (counter-balanced across change conditions), but the disfluency never preceded a target word.

Procedure Participants in the experiment were presented with 36 experimental trials and 18 filler trials. Trials were presented in semi-randomized order using a Latin-square method, such that each participant heard 18 fluent targets without and 18 disfluent targets with a filled pause. Within these groups, each participant received 6 NoChange, 6 CloseChange and 6 DistantChange trials. The presentation of the audio and visual stimuli and the recording of participants' responses was controlled by ZEP software (Veenker, 2012). During the presentation of the audio passage, a visual fixation cross appeared on the screen. When the audio had finished, the fixation cross was replaced, after a brief delay of 500 ms, by the text passage. This text passage was presented one sentence per line. Participants were instructed to indicate whether the text was a correct representation of the audio passage or an incorrect representation (i.e., the substitution of one word) by clicking with the mouse on one of two buttons labeled CORRECT and WRONG. If the participant responded that the text passage contained a substitution, he/she was asked to type the word from the audio

passage that had been replaced. The ZEP software recorded both participants' mouse-click accuracy and their mouse-click reaction times. An experimental session was self-timed and lasted approximately 40 minutes. Each experimental session finished with a post-experimental questionnaire. Participants were presented with four statements and were asked to rate their level of agreement with the statements on a scale from 1-9 (1 = strong disagreement; 9 = strong agreement). First the naturalness of the speech used in the experiment was assessed. The second question elicited accentedness ratings of the native (Experiment 1) and non-native speech (Experiment 2). Thus the 'nativeness' of both speakers, as evaluated by the listeners, could be assessed and compared across experiments. The third question assessed the impression that listeners had of the fluency of the speaker. The final question assessed the experience participants had with listening to non-native speakers of Dutch.

5.2.2 Experiment 2

Experiment 2 was identical to Experiment 1 except for the fact that now non-native speech materials were used.

Participants A sample of 40 participants participated in Experiment 2 with implicit informed consent in accordance with local and national guidelines. All were native Dutch speakers and reported to have normal hearing ($M_{age} = 24.2$, $SD_{age} = 9.2$, 4m/36f).

Design and Materials The passages used in Experiment 2 were identical to the passages from Experiment 1. For the speech materials of Experiment 2, a non-native speaker of Dutch was recorded (male, L1 Hebrew, age=45, LoR=13 years). He reported adequate knowledge of Dutch (self-reported CEFR level C1) and extensive experience with using Dutch in daily life. Recordings were made in a sound-attenuated booth using a Sennheiser ME-64 microphone (16-bit, 44000Hz, mono). The speaker was instructed to speak as clearly as possible and to make the disfluencies sound as natural as possible. In order to have a minimal contrast between the native and non-native recordings, the non-native speaker first listened to a native passage after which he was instructed to imitate the native speech. This resulted in non-native speech recordings that resembled the native recordings, but with a noticeable foreign accent. This procedure was adopted for both the experimental and the filler trials. The cross-splicing procedure was identical to Experiment 1.

Apparatus and Procedure The experimental procedure, the instructions and the post-experimental questionnaire were identical to Experiment 1.

5.3 Results

Data from both experiments were combined in all our analyses reported below.

5.3.1 Post-experimental questionnaire

Participants in both experiments had rated the naturalness, the accentedness, and the fluency of the speech stimuli on a scale from 1-9 (with higher ratings indicating more natural, more accented, more fluent speech). The average naturalness of the speech was rated 6.50, $SD = 1.78$ (native) and 4.90, $SD = 2.12$ (non-native), $t(78) = 3.65, p < 0.001$. The average accentedness of the stimuli was rated 1.23, $SD = 0.48$ (native) and 8.08, $SD = 1.10$ (non-native), $t(78) = -36.24, p < 0.001$. The fluency of the speech from both experiments was rated 5.58, $SD = 1.63$ (native) and 4.48, $SD = 1.78$ (non-native), $t(78) = 2.88, p = 0.005$. Finally, participants also rated the extent to which they regularly interacted with non-native speakers of Dutch in their daily lives: 3.03, $SD = 2.11$ (native) and 4.45, $SD = 2.32$ (non-native), $t(78) = -2.88, p = 0.005$.

5.3.2 Accuracy

The mouse click responses from Experiment 1 and 2 were analysed for participants' accuracy in responding to the change/no change question. Those trials where participants responded to have noticed a substitution, but failed to provide the correct target word, were coded as 'incorrect'. The mouse click accuracy for both experiments is illustrated in Figure 5.2. We performed mixed effects logistic regression analyses (Generalized Linear Mixed Models, GLMMs; Baayen et al., 2008; Quené & Van den Bergh, 2004, 2008) as implemented in the `lme4` library (Bates et al., 2012) in R (R Development Core Team, 2012) to analyze the mouse click accuracy from both experiments. The random part of this model consisted of the factor Participant, testing for individual differences between participants, and Item, testing for differences between items. The fixed part of the model consisted of the factor IsNonNative, testing for differences between native and non-native speech, and the factor IsDisfluent, testing for differences between fluent and disfluent trials. We also added fixed effects of the different change conditions. Because the model took the condition CloseChange as its intercept, we included the factor IsDistantChange and IsNoChange to compare the CloseChange condition with the other two conditions. Interactions between all these fixed effects were also added as predictors. This resulted in the optimal model[1] represented in Table 5.2.

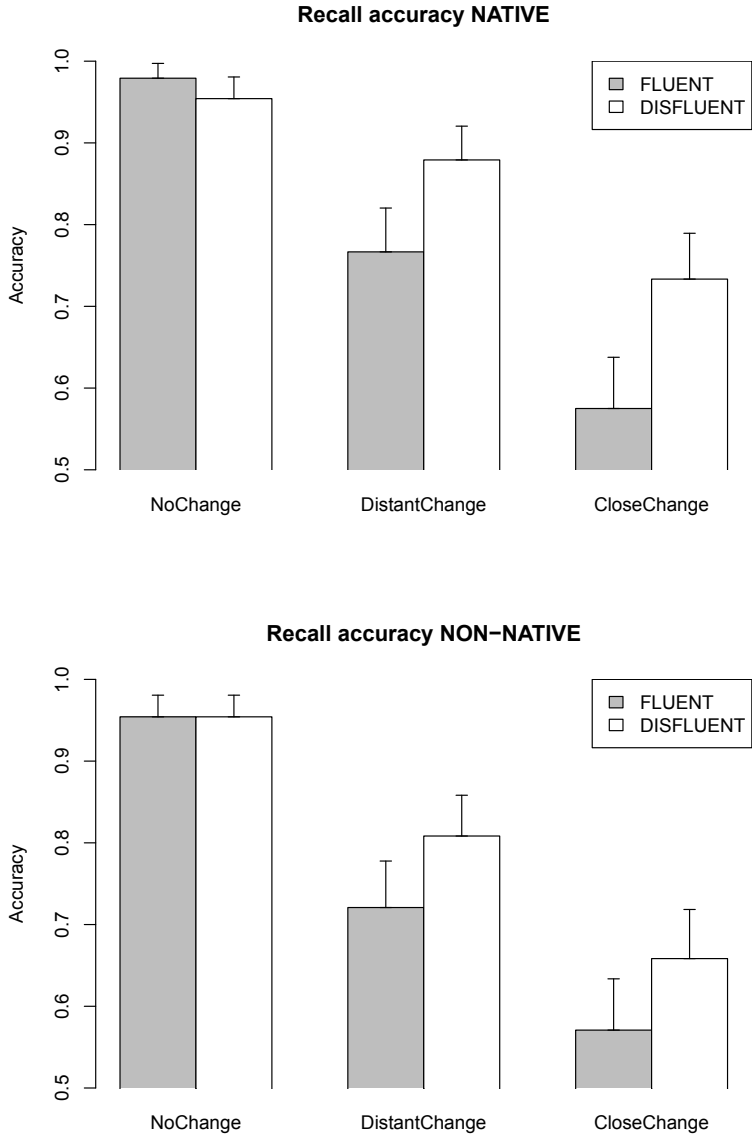
[1] We also tested a supplementary model with a maximal random part (cf. Barr, 2013; Barr et al., 2013), but this did not lead to a different interpretation of results.

Table 5.2: Estimated parameters of the mixed effects logistic regression model (standard errors in parentheses) on the mouse click accuracy in both experiments.

	estimates	z values	significance
fixed effects			
Intercept, $\gamma_{0(00)}$	0.054 (0.222)	0.24	$p = 0.808$
IsNoChange, $\gamma_{A(00)}$	4.348 (0.502)	8.66	$p < 0.001$ ***
IsDistantChange, $\gamma_{B(00)}$	0.691 (0.202)	3.43	$p < 0.001$ ***
IsDisfluent, $\gamma_{C(00)}$	0.863 (0.205)	4.21	$p < 0.001$ ***
IsDisfluent:IsNoChange, $\gamma_{D(00)}$	-1.809 (0.615)	-2.94	$p = 0.003$ **
IsDisfluent:IsDistantChange, $\gamma_{E(00)}$	-0.380 (0.297)	-1.28	$p = 0.200$
IsNonNative, $\gamma_{F(00)}$	-0.247 (0.247)	-1.00	$p = 0.318$
IsNonNative:IsNoChange, $\gamma_{G(00)}$	-0.640 (0.614)	-1.04	$p = 0.298$
IsNonNative:IsDistantChange, $\gamma_{H(00)}$	-0.459 (0.283)	-1.62	$p = 0.105$
IsNonNative:IsDisfluent, $\gamma_{I(00)}$	-0.213 (0.287)	-0.74	$p = 0.457$
IsNonNative:IsDisfluent:IsNoChange, $\gamma_{J(00)}$	1.153 (0.794)	1.45	$p = 0.147$
IsNonNative:IsDisfluent:IsDistantChange, $\gamma_{K(00)}$	0.286 (0.411)	0.70	$p = 0.487$
random effects			
Participant intercept, $\sigma^2_{u_{0(j0)}}$	0.438		
Item intercept, $\sigma^2_{v_{0(0k)}}$	0.700		

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure 5.2: Mean accuracy of mouse clicks in both experiments (error bars enclose 1.96 × SE, 95% CIs).



This model revealed that, when the transcript was identical to the previously remembered spoken passage (i.e., the NoChange condition), participants in both experiments were overall very accurate in responding that no substitution had occurred (judging from the main effect of IsNoChange, $\gamma_{A(00)}$). Also, participants in both experiments were significantly more accurate detecting substitutions involving semantically distant words compared to semantically related words (judging from the main effect of IsDistantChange, $\gamma_{B(00)}$). A beneficial effect of the filled pause was also observed (IsDisfluent, $\gamma_{C(00)}$): participants in both experiments were significantly more accurate detecting substitutions when the substitution occurred after a disfluency. However, the beneficial effect of disfluency was only found in the CloseChange and DistantChange conditions, since the interaction between IsDisfluent and IsNoChange showed that the IsDisfluency effect was attenuated in the NoChange condition. All effects reported above held for participants in both experiments, because no interactions with the factor IsNonNative were found.

5.3.3 Reaction times

Mouse reaction times (RTs) were calculated from text presentation onset onwards. Because listeners had to read through and inspect these transcripts, the RTs were relatively long (global average = 7553 ms; $SD = 4748$ ms). Statistical analyses of these data did not reveal any effect of disfluency, nativeness, or condition.

5.4 Discussion

Our experiments targeted effects of disfluencies on listeners' attention by means of the Change Detection Paradigm (CDP) which measures listeners' retention of words following disfluency. The results from our two experiments reveal that disfluencies have a beneficial effect on participants' memory. When our participants were presented with a textual representation of a just heard spoken passage, they were more accurate in detecting a change in this text when the target word in the spoken passage had been preceded by a disfluency. Thus, Experiment 1 replicates the results from Collard (2009, Experiment 3) in Dutch. These results suggest that hesitations triggered an effect on the attention directed towards the following linguistic content. In fact, the comparison between the DistantChange and CloseChange conditions demonstrated the validity of the CDP. A change involving two semantically distant words was shown to be more accurately detected than a change involving two semantically related words. This indicates that the salience of the change modulated listeners' recall, confirming that the CDP actually targets listeners' attentional mechanisms.

Experiment 2 extends the use of the CDP to the study of non-native speech. The CDP may effectively evaluate the way non-native speech induces the attention of listeners. Moreover, it allowed for a comparison between the effect of native and non-native disfluencies on attention. The beneficial effect of disfluency was found both in Experiment 1 and in Experiment 2: both native and non-native disfluencies induced heightened attention to the following content.

Our data suggest that listeners do not modulate the effect of disfluency on attention based on knowledge about the non-native identity of the speaker (cf. the results from Chapter 4, where speaker identity did affect listeners' predictive mechanisms). The reasons why listeners do not modulate the attentional effects of disfluency are unclear. It may be argued that listeners, upon encountering a disfluency, raise their attention in an automatic fashion without taking the speaker identity into account. This assumption would be in line with the Temporal Delay Hypothesis of Corley and Hartsuiker (2011): the delay inherent to both native and non-native disfluencies triggers listeners' attention.

We should, however, be careful in drawing conclusions on the basis of a null result (i.e., no interaction between IsDisfluent and IsNonNative). Several alternatives, accounting for the present null result, may be discerned. For instance, the speech materials used in the present CDPs consisted of various stories. The variation in stories may have informed the native listeners, together with the grammatical accuracy and lexical diversity, about the relatively high L2 proficiency of our non-native speaker (as assessed by LoR, L2 use, and self-reported CEFR level). Perhaps the perceived proficiency of a non-native speaker affects the way non-native disfluencies are interpreted by the listener. That is, the more proficient the non-native speaker is perceived to be, the more native-like his/her disfluencies will be interpreted. Future studies may investigate how different (perceived) levels of L2 proficiency can affect the way L2 disfluencies are processed.

Alternatively, the absence of modulation of the disfluency effect for non-native speech may result from our particular speech collection. Because we wanted to match the native and non-native as closely as possible, we used scripted passages (cf. Collard, 2009). Listeners may have been aware that our speakers acted out the speech materials, thus preventing them from interpreting the non-native disfluencies as authentically different from the native disfluencies. Future experiments, involving spontaneously produced non-native speech materials and matched native counterparts, may shed light on the generalizability of the present findings.

Despite the fact that we cannot draw definitive conclusions about how non-native disfluencies affect listeners' perceptual mechanisms, our results, nonetheless, emphasize the role of attention in an account of disfluency processing. Hesitations trigger listeners' attention with consequences for the retention of words following the hesitation.

CHAPTER 6

Conclusion

Disfluency is a pervasive feature of spontaneously produced spoken language, be it native or non-native speech. The consequences of disfluency for speech perception have been approached in the literature from two different angles, namely within the evaluative and the cognitive approach. The evaluative approach has mainly focused on non-native speech, studying for instance the (acoustic) factors that contribute to the perception of fluency. The cognitive approach has focused on native speech, investigating the effect that disfluency has on the cognitive processes involved in speech comprehension. Review of the literature reveals an apparent contradiction between, on the one hand, the *negative* effects of non-native disfluencies on subjective fluency ratings, and, on the other hand, the *positive* effects of native disfluencies on speech perception. This dissertation aimed to address this apparent contradiction by providing an account of how native and non-native fluency characteristics affect both (i) the impression that listeners have of the speaker's fluency level, and (ii) the processes involved in speech comprehension, such as prediction, memory, and attention. In this concluding chapter, the most important results of the previous chapters will be summarized, culminating in an integrative account of the negative and positive perceptual effects of disfluency. Finally, potential steps for future research will be outlined and theoretical and practical implications will be introduced.

6.1 Summary of results

6.1.1 Results of the evaluative approach to fluency

Chapters 2 and 3 adopted the evaluative approach to fluency. They investigated the effect native and non-native fluency characteristics have on perceived fluency assessment. Chapter 2 asked the question what it is that makes L2 speech sound fluent. Chapter 3 compared the way listeners assess native and non-native fluency levels.

What makes speech sound fluent? Chapter 2 investigated how raters assess the fluency levels of non-native speakers. For Experiment 1, subjective fluency judgments were collected from naïve listeners (having received specific instructions; see Appendix A). These subjective ratings were related to objective acoustic measurements of the non-native speech materials. Acoustic measurements were categorized into the three utterance fluency dimensions (Skehan, 2003, 2009; Tavakoli & Skehan, 2005): breakdown fluency (number of filled pauses, number of silent pauses, and silent pause durations), speed fluency (mean syllable duration), and repair fluency (number of corrections and number of repetitions). The acoustic measures were selected for their low intercollinearity: cross-correlations between the speech measures demonstrated that both within and across fluency aspects our speech measures were largely independent. This low intercollinearity aided the interpretation of the following analyses, in that the contribution of one fluency dimension (e.g., speed) could be separated from that of another dimension (e.g., pauses). In this fashion, we aimed to answer the first research question of Chapter 2:

> RQ 1A: What are the independent contributions of the three fluency dimensions of utterance fluency (breakdown, speed, and repair fluency) to perceived fluency?

Our results showed, first of all, that assessment of L2 fluency is largely dependent on the utterance fluency characteristics of the speech signal: the six combined acoustic measures could account for 84% of the variance of the subjective fluency judgments. Secondly, breakdown fluency and speed fluency were found to contribute most to perceived fluency ratings. In contrast, repair characteristics of the speech signal were observed to have only a weak relationship with fluency perception.

The second aim of this chapter was to seek for possible cognitive factors that underlie fluency perception. We hypothesized that differences in sensitivity to the different fluency dimensions might account for differences in correlations between acoustic measures and fluency ratings. For instance, if listeners would

be, in general, more sensitive to pause phenomena (than to speed or repair phenomena) then this could explain the large contribution of pause characteristics to the perception of fluency. A series of experiments was designed to establish the relative sensitivity of listeners to pause phenomena (Experiment 2), to the speed of delivery (Experiment 3) and to repair features in speech (Experiment 4), and thus formulate an answer to the following research question:

> RQ 1B: How well can listeners evaluate the pause, speed, and repair characteristics in speech?

The three new experiments made use of the same speech materials as Experiment 1. The participants in these three new experiments received new instructions: namely to assess the speaker's pausing behavior (Experiment 2), the speaker's speed of speaking (Experiment 3), or the speaker's use of repair strategies (Experiment 4). Subsequently, these subjective ratings were related to the objectively measured acoustic characteristics of the speech. The extent to which the objective measures accounted for the subjective judgments was taken as an indication of listeners' sensitivity to different speech characteristics. Our statistical models showed that the ratings of pausing behavior were best predicted by the objective acoustic measures. Listeners' sensitivity to the speed and repair characteristics of speech was inferior to their sensitivity to pause phenomena. This suggests that listeners are most sensitive to the pausing characteristics of speech.

Interestingly, listeners were approximately as sensitive to speed features as they were to repairs. Nevertheless, Experiment 1 had shown that repair phenomena only contribute very little to fluency judgments. The combined results from all these experiments suggest that, despite listeners' sensitivity to repair phenomena, they do not base their fluency judgments on these repair phenomena. If the perceptual sensitivity of listeners were the only factor determining the relative contributions of fluency dimensions to fluency perception, then we would expect to have found a larger contribution of repair measures to the perception of fluency in Experiment 1. Apparently, there is no direct link between listeners' perceptual sensitivity and listeners' fluency evaluation. This suggests that listeners first perceive the acoustic characteristics of a speaker's speech but then subsequently also weigh the importance of the perceived speech characteristics for fluency.

Native and non-native fluency perception Chapter 3 looked further into the weighing of acoustic fluency characteristics by comparing native and non-native fluency perception. Much of the literature on fluency assessment has focused on non-native speech (e.g., Cucchiarini et al., 2000, 2002; Derwing et al., 2004; Freed, 2000; Ginther et al., 2010; Kormos & Dénes, 2004;

Mora, 2006; Rossiter, 2009; Wennerstrom, 2000); presumably, because native speech is supposedly perceived as fluent by default. However, the psycholinguistic literature indicates that there is considerable variation in the production of disfluencies by native speakers (Bortfeld et al., 2001; Fox Tree, 1995). This raises the question:

> RQ 2: Do listeners evaluate fluency characteristics in the same way in native and non-native speech?

Because native and non-native speech differ in a large range of linguistic aspects, correlational analyses are unsuitable for comparing the perception of L1 and L2 fluency. Therefore, we applied phonetic manipulations to native and non-native speech that had been matched for one particular acoustic property. If different fluency ratings are given to two items differing in a single manipulated acoustic property, then this perceptual difference may be reliably attributed to this single manipulated acoustic property. And because the native and non-native speech has been matched, it is possible to compare the contribution of one acoustic factor across native and non-native fluency perception. Moreover, this experimental method has the additional advantage that the separate contributions of multiple acoustic factors can be investigated. The investigator can study the effect of one acoustic property on fluency judgments (e.g., the duration of pauses) whilst keeping all other possibly interacting factors (e.g., the number of pauses) constant.

Phonetic manipulations were first applied to the number and duration of silent pauses (Experiment 1), having matched the native and non-native speech materials for the number of silent pauses per 100 syllables. Three conditions were created: NoPauses - all original pauses of >250 ms had been removed; ShortPauses - all original pauses of >250 ms were altered to have a duration within the range of 250-500 ms; and LongPauses - all original pauses of >250 ms were altered to have a duration within the range of 750-1000 ms. Subjective ratings of these manipulated speech fragments from native and non-native speakers were collected in a rating experiment. Results showed that (1) native speakers were perceived to be more fluent than non-native speakers; (2) both an increase in the number of silent pauses and an increase in the duration of silent pauses negatively affected perceived fluency judgments; and (3) these manipulation effects were similar across native and non-native speech.

A similar approach was adopted in Experiment 2. Here, the speed of the speech was manipulated to compare the contribution of articulation rate and speech rate to perceived fluency. Non-native speech was increased in speed (both Articulation Rate Manipulations, ARM, and Speech Rate Manipulations, SRM) to match the mean speaking rate of the native speakers. And native speech was slowed down (both ARM and SRM) to match the mean speaking rate of the NNSs, thus making comparisons across NSs and NNSs possible.

The results from Experiment 2 mirrored those from Experiment 1. Again, (1) native speech was perceived to be more fluent than non-native speech; (2) both manipulation conditions (ARM and SRM) contributed to perceived fluency judgments: slowed down native speech resulted in lower fluency judgments, and faster non-native speech resulted in higher fluency judgments; and (3) the increase in fluency ratings of the non-native speech, and the decrease in fluency ratings of native speech, were of a similar magnitude. Based on the findings from Experiment 1 and Experiment 2, we concluded that there is no difference in the way listeners weigh the fluency characteristics of native and non-native speech. Therefore, there is no reason to believe that listeners make a qualitative distinction between native and non-native speakers when evaluating fluency.

6.1.2 Results of the cognitive approach to fluency

Chapters 2 and 3 focused on listeners' assessment of fluency. These studies demonstrated that (i) listeners weigh the perceived speech characteristics (breakdown, speed, and repair fluency) on their relevance for fluency perception, and (ii) that this weighing of acoustic factors is similar for native and non-native fluency assessment. These observations do not necessarily warrant the conclusion that native and non-native disfluencies are perceptually equivalent, because Chapters 2 and 3 have only investigated the effects of disfluencies on listeners' subjective impressions of the speaker. Chapters 4 and 5 adopted the cognitive approach to fluency to test whether native and non-native disfluencies have different effects on the cognitive processes involved in speech comprehension. Chapter 4 asked the question whether native and non-native *uhm*'s may guide prediction of low-frequency referents to the same extent. Chapter 5 compared the effects of native and non-native disfluencies on attention.

Disfluency and prediction The psycholinguistic literature on disfluencies in native speech seems to converge on the conclusion that native disfluencies may aid the listener in comprehension. Listeners use their experience with the regularities in the distribution of disfluencies to anticipate the linguistic content following a disfluency. The literature on disfluency production indicates that disfluencies tend to occur before open-class words (Maclay & Osgood, 1959), unpredictable lexical items (Beattie & Butterworth, 1979), low-frequency color names (Levelt, 1983), or names of low-codability images (Hartsuiker & Notebaert, 2010). Therefore, disfluencies cue the onset of dispreferred or more complex content.

These conclusions have been drawn based on studies of disfluent native speech. It is unknown how disfluencies in non-native speech may affect listeners' predictive strategies. Therefore, Chapter 4 compared the way native and non-native disfluencies affect listeners' predictive strategies. We hypothesized that,

due to the fact that there are less regularities in the distribution of non-native disfluencies, non-native disfluencies would be worse predictors of the word to follow (as compared to native disfluencies).

Previous literature investigating disfluency effects on prediction have reported that listeners may interpret native disfluency as a symptom of speaker difficulty in conceptualization. For instance, listeners can attribute disfluency to trouble in recognizing unknown objects (e.g., Arnold et al., 2007) or to trouble with the object's discourse status (e.g., Barr & Seyfeddinipur, 2010). In order to compare native and non-native disfluency, we targeted listeners' attribution of disfluency to difficulty in formulation (i.e., trouble in lexical access, rather than in conceptualization). We argued that it is at this particular stage in speech planning that native and non-native speakers diverge.

Therefore, the first research question of Chapter 4 read:

> RQ 3A: Do listeners anticipate low-frequency referents upon encountering a disfluency?

This question was addressed by means of eye-tracking experiments. An adapted version of the methodology of Arnold et al. (2007) was used: participants were presented with pictures of high-frequency (e.g., a hand) and low-frequency objects (e.g., a sewing machine). Simultaneously, fluent and disfluent spoken instructions were played (e.g., 'Click on the red [target]' vs. 'Click on *uh* the red [target]'). It was hypothesized that listeners might attribute the presence of the disfluency to speaker difficulty in formulating the label for the low-frequency object (rather than the high-frequency object). This would result in more looks to the low-frequency object, prior to target onset, when listeners heard native disfluent speech.

Experiment 1 failed to provide evidence for native disfluencies affecting listeners' predictive strategies. Two possible factors were identified that might have been responsible for this null result, namely (1) the presence of a picture familiarization phase (prior to the eye-tracking experiment), and (2) the long time span between the disfluency *uh* and target onset (i.e., the presence of color adjectives preceding the target). Therefore, Experiment 2 involved a new version of Experiment 1, without a familiarization phase and with shorter speech stimuli (i.e., without color adjectives: 'Click on the [target]' vs. 'Click on *uh* the [target]').

The results from Experiment 2 indicated that listeners had a preference, prior to target onset, for looking towards the low-frequency object (e.g., sewing machine) as opposed to the high-frequency object (e.g., the hand). This preference was only observed in the disfluent condition, not in the fluent condition: only when hearing disfluent speech did listeners anticipate reference to a low-frequency object. These results suggest that listeners attribute the presence of disfluency to speaker difficulty in formulation.

The third experiment was designed to allow for a comparison of the effects of native and non-native disfluencies on prediction:

> RQ 3B: Do native and non-native disfluencies elicit anticipation of low-frequency referents to the same extent?

Experiment 3 was identical to Experiment 2, but the participants in Experiment 3 listened to non-native speech. The results from Experiment 2 and 3 were combined to test for an interaction between the type of speaker (native vs. non-native) and the presence of a disfluency bias for low-frequency referents. It was found that, where native disfluencies elicited anticipation of low-frequency referents, non-native disfluencies did not. We argue that listeners reduced their use of disfluencies for prediction when listening to an L2 speaker, because non-native disfluencies are worse predictors of the linguistic content to follow. These results suggest that knowledge of the non-native identity of a speaker, as evidenced by a foreign accent, influences the way listeners use performance aspects of the speech signal (i.e., disfluency) to guide prediction.

Disfluency and attention Where in Chapter 4 disfluency effects on prediction were targeted, Chapter 5 studied how native and non-native disfluencies affect listeners' attention. Earlier work on perception effects of disfluencies showed that native disfluencies may trigger listeners' attention. This raises the question whether there are differential effects of native and non-native disfluencies on attention, as addressed by the research question of Chapter 5:

> RQ 4: Do native and non-native disfluencies trigger heightened attention to the same extent?

Disfluency effects on listeners' attention could be the result of the non-arbitrary distribution of native disfluencies: disfluencies cue relatively more complex information and, therefore, listeners may benefit from raised attention levels in order to ensure timely comprehension of the complex content. The distribution of disfluencies in non-native speech is, from the native listener's point of view, more irregular than the disfluency distribution in native speech. Therefore, non-native disfluencies are worse cues of upcoming, relatively more complex information. We hypothesized that the effect of non-native disfluencies on listeners' attention might therefore be attenuated (relative to that of native disfluencies; cf. Chapter 4).

Alternatively, disfluency effects on listeners' attention could also be the result of more automatic cognitive processes in response to delay. The *Temporal Delay Hypothesis* (Corley & Hartsuiker, 2011) argues that the temporal delay that is inherent to disfluency facilitates listeners' recognition and listeners' retention of words. Thus, both native and non-native disfluencies would trigger heightened attention levels.

We investigated attentional effects of native and non-native disfluencies by means of the Change Detection Paradigm (CDP). Participants were instructed to remember a spoken passage of three sentences. One of the words (the 'target') in the passage was either presented in a fluent or disfluent context ("... He saw that the patient with the [*uh*] *wound* ..."). After listening to the spoken passage, participants were presented with a textual representation of the memorized spoken passage. This text sometimes contained a substitution of the target word. Participants were asked to indicate whether the text was a correct representation of the spoken passage or whether the text contained a substitution. We hypothesized that, if disfluencies trigger listeners' attention, then participants should be better in detecting a change to a target word that had been presented in a disfluent context (e.g., the *uh* wound) relative to the same target word in a fluent context (e.g., the wound).

We designed two experiments: participants in Experiment 1 were presented with fluent and disfluent native speech, and participants in Experiment 2 heard non-native speech. The results from both experiments indicated that disfluency had a beneficial effect on participants' recall: listeners were more likely to detect a substitution of a word that had been preceded by a disfluency than substitutions of words in fluent context. This disfluency effect was present in both experiments: both native and non-native disfluencies triggered heightened listeners' attention. No attenuation of the disfluency effect was observed when participants listened to non-native speech.

These findings suggest that listeners do not modulate the effect of disfluency on attention based on knowledge about the non-native identity of the speaker. This could indicate that listeners, upon encountering a disfluency, raise their attention in an automatic fashion without taking the speaker identity into account (supporting the Temporal Delay Hypothesis of Corley & Hartsuiker, 2011). However, several concerns were raised about the methodology reported in Chapter 5 (e.g., the scripted nature of the speech, the perceived L2 proficiency of the non-native speaker of Chapter 5). Therefore, we should be careful in drawing conclusions about attentional effects of non-native disfluencies on the basis of a null result (i.e., no interaction between the disfluency effect and the type of speaker [native vs. non-native]). Despite the fact that we cannot draw definitive conclusions about how non-native disfluencies affect listeners' perceptual mechanisms, our results, nonetheless, emphasize the role of attention in an account of disfluency processing. Hesitations trigger listeners' attention with consequences for the retention of words following the hesitation.

6.2 An integrative account of fluency perception

This dissertation addressed the following main research question:

> Main RQ: How do fluency characteristics affect the perception of native and non-native speech?

This research question was motivated by an apparent contradiction between, on the one hand, the *negative* effects of non-native disfluencies on subjective fluency ratings, and, on the other hand, the *positive* effects of native disfluencies on speech perception. The combined results from the previous chapters contribute to our understanding of the beneficial, and of the disadvantageous effects of native and non-native disfluency on listeners. In the following paragraphs, an attempt will be made to demonstrate that the results from this dissertation can resolve the apparent contradiction by providing answers to the main research question.

6.2.1 Listeners' subjective impressions

Following the framework by Segalowitz (2010), we have argued that utterance fluency characteristics follow from the speaker's cognitive fluency. Disfluency is a symptom of inefficiency of the processes involved in speech planning and production. This inefficiency may arise at any of the stages in speech production: in finding out what to say (conceptualization), in finding the right words (formulation), or in generating a phonetic plan (articulation; Levelt, 1989). Both native and non-native speakers suffer from disfluency, because both types of speakers experience the time pressure under which natural conversations take place.

This does not mean that native and non-native disfluency production are identical. There are considerable quantitative and qualitative differences between native and non-native disfluency production. Regarding the quantitative differences, non-native speakers produce more disfluency: L2 cognitive fluency is less efficient than L1 cognitive fluency. With respect to the qualitative differences, non-native speakers produce disfluencies at different points in the utterance: inefficiency in L2 cognitive fluency has different origins compared to L1 cognitive fluency (De Bot, 1992; Segalowitz, 2010). Insufficient declarative (knowledge) and procedural (skill) mastery of the L2 have been identified as two sources of L2-specific inefficiency (cf. De Jong et al., 2012a).

The quantitative differences between native and non-native cognitive fluency (which surfaces in utterance fluency as a difference in the number of disfluencies) means that, in practice, non-native speakers are generally perceived

to be less fluent than native speakers. The qualitative difference between native and non-native cognitive fluency (surfacing in a different disfluency distribution) does not seem to affect the way native and non-native fluency characteristics are weighed (e.g., a native speaker pausing before a low-frequency word is 'just as bad' as a non-native speaker pausing before a high-frequency word). Apparently, both native and non-native fluency characteristics are perceived to be symptoms of reduced cognitive fluency and, therefore, listeners weigh the fluency characteristics of native and non-native speech in a similar way (cf. Chapter 3).

6.2.2 Listeners' predictive strategies

However, listeners are not insensitive to the qualitative differences between native and non-native cognitive fluency. Chapter 4 has demonstrated that listeners can make use of symptoms of cognitive inefficiency by using disfluency as a cue to upcoming, relatively more complex information. Listeners were only observed to use disfluency to predict reference to low-frequency objects when listening to a native speaker. When listeners heard a non-native speaker produce similar spoken instructions, they did not use disfluency to guide anticipation of low-frequency referents. This suggests that the non-native identity of the speaker can modulate the effect that disfluencies have on prediction.

These findings from Chapter 4 suggest that listeners are sensitive to the qualitative differences between native and non-native cognitive fluency. Listeners are familiar with the regularities in native disfluency production and, therefore, can use disfluency to anticipate relatively more complex information. They are also familiar with the more irregular distribution of non-native disfluencies and, therefore, attenuate the effect of non-native disfluencies on prediction. As such, listeners can, in a very clever way, make use of symptoms of inefficiency for prediction.

6.2.3 Listeners' attentional resources

Chapter 5 reported that both native and non-native disfluencies have beneficial effects on speech comprehension because they were observed to trigger listeners' attention. Thus, the positive effect of native disfluencies on attention, reported in the literature, is extended to non-native disfluency. Despite several concerns about the methodology of the experiments in Chapter 5, we may yet speculate as to the reasons why native and non-native disfluencies heighten listeners' attention.

One possible explanation of the attentional effects of disfluency is related to the Temporal Delay Hypothesis (proposed by Corley & Hartsuiker, 2011). This hypothesis argues that temporal delay in the speech signal improves speech

comprehension. The temporal delay may provide the listener with additional time to orient to the upcoming information (disengagement from previous linguistic content and shift to new information). This would suggest that disfluency effects on attention are an automatic consequence of intrinsic temporal delay, independent of knowledge about the identity of the speaker. Both native and non-native disfluencies inherently introduce temporal delay and, therefore, both types of disfluencies trigger an orienting response.

There are, however, some empirical findings that challenge the Temporal Delay Hypothesis. One of its conjectures is that any kind of delay in the speech signal should improve speech comprehension: disfluencies, such as filled pauses and silent pauses, but also coughs, beeps, or barks. However, the evidence from previous studies for beneficial effects of delays (that are not disfluencies) on speech comprehension is equivocal (cf. Bailey & Ferreira, 2003; Barr & Seyfeddinipur, 2010; Corley & Hartsuiker, 2011). For instance, Fraundorf and Watson (2011) found that filled pauses did improve listeners' recall of previously remembered stories, but coughs, matched in duration to the filled pauses, did not.

Alternatively, the null result in Chapter 5 may be explained by listener strategies in response to anticipated comprehension effort. Both native and non-native disfluencies arise through relatively high cognitive load in speech production. This cognitive load experienced by the speaker may also carry consequences for listener effort in speech comprehension. For instance, finding the right label for a low-frequency object may present a native speaker with additional cognitive load, as evidenced by a higher probability of disfluency. At the same time, low-frequency words are also more cognitively demanding (for the listener) to comprehend, as evidenced by slower responses in word recognition (e.g., Marslen-Wilson, 1987) and lower recognition accuracy (e.g., Goldinger, Luce, & Pisoni, 1989). Listeners may anticipate this increased effort upon encountering a disfluency and, therefore, adopt precautionary comprehension strategies, such as the raising of attention.

The strategy of heightened attention levels in response to disfluency may also apply to the processes involved in comprehension of non-native speech. For instance, the difficulty experienced by a non-native speaker in planning and producing L2 speech may result in semantic inaccuracy, grammatical errors, or poor pronunciation. All of these challenge the listener in comprehension. Therefore, listeners may benefit from strategically raising their attention levels when encountering disfluency - both when listening to native and non-native speech - to ensure timely comprehension of cognitively demanding linguistic input.

Based on the data from Chapter 5, we cannot yet discriminate between these two explanations of the attentional effects of native and non-native disfluencies. New investigations will have to determine the value of either of the two

explanations (see next paragraph). Nevertheless, the combined findings from the different chapters in this dissertation do resolve the apparent contradiction between, on the one hand, the *negative* effects of non-native disfluencies on subjective fluency ratings, and, on the other hand, the *positive* effects of native disfluencies on speech perception.

We argue that negative and positive effects of disfluency are the result of different listener considerations. Native and non-native disfluencies have negative effects on listeners' judgments about the speaker's fluency level, because listeners are assumed to consider both types of disfluencies to be symptoms of speech production difficulty. Despite these negative effects, listeners are capable of using these symptoms of speaker difficulty to anticipate reference to relatively more complex linguistic content (e.g., low-frequency words). However, listeners only adopt this predictive strategy when listening to native speech, because of the regularities in the distribution of native disfluencies. The distribution of non-native disfluencies is, from the native listener's point of view, much more irregular, leading to an attenuation of the effects of non-native disfluency on prediction. With respect to attention, both native and non-native disfluencies may heighten listeners' attention to the following information, either because of the delay intrinsic to native and non-native disfluency, or because of listeners taking precautionary measures to reduce anticipated cognitive effort in comprehension.

6.3 Future research

Our integrative account of the perception of fluency may motivate future studies to test the account's conjectures and/or to expose its limitations. For instance, our conclusions about perceived fluency (drawn in Chapters 2-3) were based on experiments in which we presented raters with specific instructions on how to assess fluency. Current language tests commonly provide their raters with explicit instructions about how to assess oral fluency by reference to utterance fluency characteristics, such as speed of delivery, pauses, and hesitations. This tendency is found both within language testing practice and within the literature on fluency perception (e.g., Derwing et al., 2004; Rossiter, 2009). Because we also adopted this procedure in Chapters 2-3, our findings could be directly applied to language testing practice where similar methods are used.

However, exactly because of the prevalence of very specific fluency instructions, the relationship between ratings of fluency in the broad sense (i.e., overall speaking proficiency) and ratings of fluency in the narrow sense (i.e., a component of overall speaking proficiency) has not (yet) received much attention. Some studies have investigated the componential nature of overall speaking proficiency (Adams, 1980; Higgs & Clifford, 1982; McNamara, 1990), or

have targeted the factors that contribute to oral proficiency (Ginther et al., 2010; Iwashita et al., 2008; Kang et al., 2010). However, an investigation of how fluency characteristics in the speech signal contribute to ratings of fluency in the narrow and in the broad sense has not yet been undertaken.

The dearth in studies comparing the narrow and broad sense of fluency carries consequences for our conclusions in Chapter 3. The conclusion that native and non-native fluency perception are comparable, was drawn on the basis of data collected through evaluations of fluency in the narrow sense. This raises the question whether the similarity of native and non-native fluency perception also applies when listeners assess fluency in the broad sense (for instance, in fluency assessment without any instructions on what comprises fluency). This question is very much relevant for everyday situations in which interlocutors in a conversation draw inferences about the other (native or non-native) speaker's social status (Brown et al., 1975), emotion (Scherer, 2003), physical properties (Krauss et al., 2002), metacognitive state (Brennan & Williams, 1995), fluency level (e.g., Chapter 2 and 3), etc. Listeners' considerations in these spontaneous, uncontrolled situations have been under-investigated in the literature and future studies may find ways of tapping listeners' underlying deliberations in these situations. Until that time, it is uncertain whether our conclusions about native and non-native fluency in the narrow sense generalize to situations without clearly formulated fluency assessment instructions.

Another issue that prospective studies may address is related to the differential effects of native and non-native disfluencies on prediction. Chapter 4 has revealed that non-native disfluencies do not guide prediction of low-frequency referents (whereas native disfluencies do). This does not necessarily imply that non-native disfluencies do not guide prediction at all. Our experiments targeted listeners' attributions of disfluency to speaker trouble in *formulation*. It is, as yet, unclear whether non-native disfluencies also have differential effects on speech comprehension (relative to native disfluencies) when listeners attribute disfluency to speaker trouble in *conceptualization*. For instance, an experiment may be proposed in which, following Arnold et al. (2007), listeners are presented with visual arrays of known vs. unknown objects (e.g., a picture of an ice-cream cone paired to a picture of an abstract symbol). When a native speaker is heard producing disfluent instructions to click on one of the objects, listeners have been shown to anticipate reference to the unknown object (Arnold et al., 2007), suggesting that listeners attribute the disfluency to trouble in conceptualization of the unknown object. A new experiment may test how listeners deal with a situation in which a non-native speaker struggles to produce these kinds of instructions.

It could be argued that non-native disfluency does not affect listeners' predictive mechanisms in any situation. Listeners may consider the non-native speaker to have equal trouble with the production of known and unknown words

in their L2 (i.e., naming an ice-cream cone is more difficult for an L2 speaker as compared to an L1 speaker). In this case, the perception of non-native speech would pattern with the perception of atypical native speakers (e.g., a native speaker with object agnosia; Arnold et al., 2007). Alternatively, one could also hypothesize that non-native disfluency, just like native disfluency, may guide prediction of unknown referents. This would suggest that listeners are aware that both native and non-native speakers encounter similar troubles in conceptualizing unknown referents (in contrast to L2-specific difficulty in formulating high-frequency referents). As such, the non-native identity of the speaker would play a role in listeners' attributions of disfluency to speaker trouble at the phase of conceptualization. In this fashion, new studies into prediction as a component of speech comprehension may determine where the non-native identity of a speaker plays a role in speech comprehension, and where it does not.

A final issue that may encourage follow-up studies concerns the two introduced explanations for the attentional effects observed in Chapter 5: they are either due to the delay intrinsic to native and non-native disfluency (the Temporal Delay Hypothesis; Corley & Hartsuiker, 2011), or they are the result of listeners taking precautionary measures to reduce anticipated cognitive effort in comprehension. New experiments may be designed that test these two explanations. For instance, researchers may present listeners with speech that contains forms of delay that do not necessarily cue more complex linguistic content (e.g., coughs). If listeners are found to be more accurate in recalling words that were preceded by such delays, this would suggest that delay alone can account for heightened attention. Thus, the field of speech perception may benefit from investigations into the effects of different types of delay in various comprehension tasks (e.g., recognition, prediction, retention, syntactic parsing, etc.).

Since the definitive explanation for the observed attentional effects of disfluency is, as yet, lacking, the relationship between the attentional and prediction effects of disfluency is also unclear. If both attentional and prediction effects of disfluency are caused by listeners' experience with the regularities of disfluency production, another interesting question regards the time course of these two types of effects. Does prediction precede heightened attention? Or vice versa? And does one effect elicit the other? Do heightened attention levels trigger predictive mechanisms, or does prediction of relatively more complex information implicate the attention levels required for the processing of this information? These questions may form an agenda for future investigations into the cognitive effects of disfluency.

6.4 Implications

The account of fluency perception, introduced above, has proposed that listeners view disfluency in spoken language as a symptom of inefficiency in speech production. These symptoms have a negative effect on fluency judgments, but they can also have positive effects on cognitive processes involved in speech comprehension, such as prediction and attention. The proposed account carries implications for both the evaluative and cognitive approach to fluency.

6.4.1 Implications for the evaluative approach to fluency

Rater instructions The findings from Chapter 2 have revealed that the pausing and speed characteristics of L2 speech are the most important contributors to perceived fluency judgments. Non-native speakers' repair strategies (e.g., corrections, repetitions) also contribute to fluency perception, but these acoustic features play a much smaller role. These observations are applicable to testing procedures in language testing practice. Many language tests use speaking rubrics with explicit mention of aspects of fluency (e.g., TOEFL iBT, ACTFL, IELTS), but the way in which raters are instructed to assess fluency differs. For instance, for IELTS (IELTS, [n.d.]), raters are instructed to assess, amongst others, speakers' 'fluency and coherence' on a 9-band proficiency scale. Descriptives of speech at each of the 9 proficiency bands are provided, such as references to the length of the speech performance, pauses, hesitations, repetitions, and self-corrections. The descriptives of several bands contain reference to repair strategies: for example, speakers at band 5 use "repetition, self-correction and/or slow speech"; and speakers at band 6 are described as "willing to speak at length, though [they] may lose coherence at times due to occasional repetition, self-correction or hesitation". In contrast, hardly any descriptives contain reference to speed characteristics. The rating procedure of this language test (and also others) may be informed by our hierarchy of fluency dimensions, as described in Chapter 2. For instance, the contributions of pause and speed characteristics can be stressed, whereas reference to repair strategies could be minimized. This does not mean that references to repair strategies should be removed from rater instructions, but they certainly should not be emphasized either.

In a similar way, our findings about the relevance of the different utterance fluency dimensions for fluency perception can also be applied to instruments for automatic fluency assessment. Such instruments are already being used in official language tests, such as the PTE Academic, TOEFL iBT, and the Dutch Immigration Test. Our results can guide designers of these tests to adjust the weights that are applied to automatically measured acoustic fluency characteristics. For instance, it is possible to implement a higher fluency penalty on

pausing phenomena as compared to repair phenomena. In this fashion, automatic fluency assessment is expected become a better approximation of human fluency judgments.

Relevance of fluency dimensions The first experiment from Chapter 2 suggested a hierarchy in the contributions of fluency dimensions to fluency perception: pause and speed characteristics of speech contribute most to perceived fluency judgments, and repair strategies contribute only very little. The following three experiments of Chapter 2 addressed the question why listeners adopt this hierarchy of fluency dimensions. We proposed that differences in perceptual sensitivity to the three fluency dimensions might account for this result. If pause and speed characteristics are more salient than repair phenomena, then this may explain why raters base their fluency judgments more on the speaker's pause behaviour and speed of speaking rather than on the speaker's repair behaviour. However, listeners were found to be as sensitive to repair characteristics as to speed characteristics. Therefore, perceptual sensitivity alone cannot account for the observed relative contributions of the different fluency dimensions to fluency perception.

What potential other factor may account for the observed hierarchy of fluency dimensions? Why do listeners consider a speaker's pause behavior to be a better indicator of the speaker's fluency level as compared to the speaker's repair behavior? One might hypothesize that pauses are better indicators of an L2 speaker's overall speaking proficiency (relative to repairs and repetitions). De Jong et al. (2012b) investigated the relationship between measures of L2 utterance fluency and measures of L2 cognitive fluency, by collecting data from a large cohort of non-native speakers ($N = 179$). Utterance fluency was operationalized as a set of acoustic fluency characteristics (e.g., articulation rate, number of silent pauses, number of corrections, etc.). Cognitive fluency was operationalized as the sum of the speakers' L2 linguistic knowledge and processing skills (e.g., L2 grammatical knowledge, speed of L2 lexical retrieval, etc.). Relating the utterance fluency measures to the cognitive fluency measures, the authors found that a speaker's L2 proficiency correlated most strongly with the speaker's (inverse) articulation rate: 50% of individual variance in (inverse) articulation rate was explained by the speaker's L2 cognitive fluency. In contrast, average pausing duration was only weakly related to linguistic knowledge and processing skills. This finding suggests that inefficiency in L2 speech production primarily surfaces in the non-native speaker's articulation rate, rather than his/her average pausing behavior.

Nevertheless, the literature on fluency perception has repeatedly argued that listeners do take a speaker's pausing behavior to be indicative of that person's fluency level (Cucchiarini et al., 2002; Derwing et al., 2004; Rossiter,

2009). This presents an apparent conundrum to the study of fluency production and perception: listeners base their fluency judgments on pausing and speed characteristics, while only speed of articulation truly reflects the L2 speaker's underlying cognitive fluency. One possible solution to the puzzle may lie in the perceptual relationship between speakers' speed of speaking and speakers' pausing behavior. Both the study of De Jong et al. (2012b) and the experiments in Chapter 2 used 'independent' acoustic measures: that is, pause and speed measures that portray low intercollinearity, such as the measure *articulation rate* which is calculated by a speaker's *speaking time*, excluding silent pauses (vs. the measure *speech rate* which is calculated by a speaker's *total time*, including silent pauses). This approach is useful when one wants to distinguish the separate contributions of the three fluency dimensions. However, we do not know whether listeners are also capable of perceptually distinguishing speed fluency from breakdown fluency. Alternatively, one could hypothesize that these two dimensions together load onto one perceptual category: pause&speed fluency.

Further inspection of the data from Chapter 2 supports this latter suggestion. In Chapter 2, Experiment 2 was designed to collect perceptual judgments of L2 speakers' pausing behavior, and Experiment 3 was designed to collect perceptual judgments of L2 speakers' speed of speaking. The correlations between objective pause measures and the speed measure, as reported in Table 2.3, did not exceed $r = 0.4$. Nevertheless, supplementary analysis of the relationship between the subjective judgments of pausing behavior (Experiment 2) and speed behavior (Experiment 3) reveals a strong correlation between the pause and speed ratings: Pearson's $r = 0.839, p < 0.001$. This indeed suggests that pause and speed characteristics, despite weak correlations in utterance fluency, do load onto one pause&speed percept in perceived fluency. Both independent parts of this percept are strong predictors of perceived fluency judgments (Cucchiarini et al., 2002; Derwing et al., 2004; Rossiter, 2009, and Chapter 2 of this dissertation), but only an L2 speaker's speed of articulation - not pausing - is strongly correlated with L2 knowledge and skills (i.e., oral proficiency; De Jong et al., 2012b).

This raises the question which fluency construct language tests should reflect: perceived fluency or cognitive fluency? The current situation in language testing practice is that language test scores approximate perceived fluency judgments. This is inherent to many language tests because they make use of human raters. Thus, a low score on a language test correlates with low subjective fluency judgments. However, instead of reflecting perceived fluency, language tests could also reflect the underlying efficiency of the cognitive processes involved in L2 speech production, that is, cognitive fluency. Language tests evaluating cognitive fluency would provide insight into the speaker's underlying L2 knowledge and skills. As such, these tests require automatic assessment of those speech characteristics that reflect an L2 speaker's underlying L2 proficiency (e.g., ar-

ticulation rate; De Jong et al., 2012b), forgoing human perception.

Ultimately, designers of language tests will have to decide what fluency construct their language test is to reflect, on the basis of the particular goals of the designers. The outcomes of language tests that have been designed to reflect native listeners' impressions (perceived fluency) are applicable to social interactions with native speakers. Scores on such language tests inform non-native speakers about how native speakers will perceive their proficiency. Alternatively, if a language test is designed to reflect underlying L2 proficiency (cognitive fluency), its outcomes provide insight into speakers' underlying L2 knowledge and skills (irrespective of native speakers' subjective prejudices, stereotypes, etc.). Thus, language tests reflecting cognitive fluency constitute are expected to be more objective assessment tools, since they produce highly reliable output.

Norms and standards On the basis of the results from Chapter 3, we concluded that, when raters are instructed to evaluate fluency in the narrow sense, native and non-native fluency perception are comparable. This implies that listeners do not make a qualitative distinction between native and non-native speakers in fluency assessment. Rather, the difference between native and non-native fluency is gradient: variation in fluency judgments between different native and non-native speakers can be accounted for by quantitative differences. This conclusion is good news for language learners. Our results suggest that there is no insurmountable obstacle preventing them from achieving native-like fluency levels in their second language.

The conclusion that native and non-native fluency perception are comparable also entails that, contrary to common opinion, native speech is not perceived to be fluent by default; instead, there is considerable variation in the perceived fluency of native speakers. Variation in fluency characteristics of native speech (artificially created through phonetic manipulations (Chapter 3) or naturally observed variation; Bortfeld et al., 2001) influences how fluent native speakers are perceived to be. This observation led us to conclude, in Chapter 3, that a single ideal native fluency standard does not exist.

This carries consequences for language testing practice, where non-native fluency levels are regularly assessed on grounds of an idealized disfluency-free norm. However, the results from Chapters 4 and 5 show that disfluency-free speech should not be considered the norm, because disfluencies can serve a purpose in speech comprehension. They may guide listeners to predict upcoming content and/or may trigger heightened attention levels. This suggests that, instead of penalizing all disfluent speech, language tests should rather aim to assess fluency by penalizing only those speech characteristics that hinder communication.

6.4.2 Implications for the cognitive approach to fluency

The observations reported in Chapter 4 contribute to our understanding of the predictive strategies of listeners. More specifically, they add to our knowledge of (1) the content of the predictions, (2) the factors that form the basis of prediction, and (3) the factors that modulate prediction strategies.

Our results inform scholars in the field of speech perception about the content of listener predictions. Previous studies had already demonstrated that disfluency triggers prediction of discourse-new (Arnold et al., 2004; Barr & Seyfeddinipur, 2010) and unknown objects (Arnold et al., 2007; Watanabe et al., 2008). This involved listeners attributing disfluency to speaker trouble in conceptualization. The experiments reported in Chapter 4 demonstrate that listeners can also flexibly attribute disfluency to speaker trouble in formulation (i.e., lexical retrieval of low-frequency words). Apparently, listeners are adept at attributing symptoms of speech production inefficiency to different stages in speech production.

Our results also show that prediction is not merely based on linguistic phenomena (such as the semantic value of verbs, the syntactic characteristics of sentences, or the phonological properties of words; Altmann & Kamide, 1999; DeLong et al., 2005; Van Berkum et al., 2005), but also on the performance of an utterance (cf. Arnold et al., 2007; Barr & Seyfeddinipur, 2010; Dahan et al., 2002; Weber et al., 2006). Listeners take into account all sorts of information that is available in the speech signal to achieve an appropriate understanding of spoken language. This insight may inform researchers, but also communication specialists and public speakers, that successful communication is not only dependent on *what* is said, but also on *how* it is said. Not only the content of spoken language, but also the speaker's performance contributes to communication. This insight may encourage scholars to identify other factors, external to the linguistic content of the utterance, that contribute to prediction and speech comprehension in general, such as emotion, indirect meaning, stereotypes, etc.

Finally, Chapter 4 also identifies one possible factor that can modulate prediction, namely knowledge about the non-native identity of the speaker. Apparently, speech comprehension is not only dependent on *what* is said, or on *how* it is said, but also on *who* says it. Previous research has reported on listeners adjusting their perceptual mechanisms in response to hearing a foreign accent (e.g., attenuation of the P600 effect; Hanulíková et al., 2012). This research, supplemented by our findings on prediction, reveals that comprehending non-native speech is different from comprehending native speech.

6.5 Conclusions

This dissertation has proposed an account of fluency perception that argues that disfluency arises from speech production difficulty. The symptoms of such speaker trouble (i.e., individual disfluencies in the speech signal) negatively affect listeners' impressions of the speaker's fluency level. Nonetheless, listeners can, in a very clever way, make use of the disfluent character of spontaneous speech for comprehension: disfluencies may elicit anticipation of - and heightened attention to - subsequent linguistic content. This account of fluency perception is the result of combining the evaluative approach and the cognitive approach to fluency. We hope to have convinced the reader of the advantages of combining these different approaches in order to arrive at a more comprehensive account of how native and non-native fluency characteristics are perceived. More specifically, because of the combination of the evaluative and the cognitive approach, the findings from one approach can inform the other.

For instance, the evaluative approach can help put the findings from the cognitive approach into perspective. Many studies into the cognitive effects of disfluency report on experiments that use monologue speaking conditions, involving artificially manipulated, one-sentence stimuli. However, spoken communication in everyday life takes place in real face-to-face conversations with free interaction between multiple interlocutors. In these natural circumstances, social factors play a major role. For instance, in natural conversations people do not take the speech signal merely as input for comprehension but also as input for social inferencing (e.g., the speech signal as an indicator of social status, physical attributes, and cognitive state). The evaluative approach to fluency reminds the cognitive approach that fluency perception, ultimately, takes place in a social setting outside of the psycholinguistic lab. In these natural communicative settings, disfluency does not only guide prediction or attract listeners' attention, but it also forms an integral part of social interaction in which people form subjective impressions of the other persons around them. Without an understanding of how listeners arrive at these socially relevant impressions, the findings from the cognitive approach are limited to very specific, highly controlled, communicative situations.

Conversely, the findings from the cognitive approach also inform the evaluative approach to fluency. Chapters 4 and 5 advocate a more fine-grained perspective on fluency assessment. These chapters demonstrate that disfluencies can sometimes help the listener in comprehension. Therefore, fluency assessment should not take place along disfluency-free standards, but language tests should ultimately discriminate between those speech characteristics that help, and those that hinder communication. This calls for a large-scale research program that aims to provide a way of distinguishing communicatively 'helpful'

and 'unhelpful' speech characteristics. It will have to combine methods and insights from various scientific disciplines, such as (applied) linguistics, communication sciences, psychology, and their various subfields, in order to fully understand the mechanisms involved in human spoken communication.

This dissertation has contributed to the study of spoken communication through an investigation of fluency, which combined the evaluative and the cognitive approach. Thus, it has extended our understanding of how both native and non-native fluency characteristics are weighed for fluency assessment, and how these characteristics affect speech comprehension. The described implications of our studies and the call for new investigations testify to the notion that speech performance matters: communication through spoken language does not only depend on what is said, but also on how it is said and by whom. As such, the studies in this dissertation have put speech performance at the heart of the study of spoken communication, right where it belongs.

References

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Frith (Ed.), *Measuring spoken language proficiency* (p. 1-6). Washington, DC: Georgetown University Press.

Almeida, J., Knobel, M., Finkbeiner, M., & Caramazza, A. (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review*, *14*(6), 1177-1182.

Altmann, G. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, *137*(2), 190-200.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*(1), 25-36.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say -thee uh- you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914-930.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578-582.

Arnold, J. E., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 28-55.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*(2), 183-200.

Barr, D. J. (2008a). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457-474.

Barr, D. J. (2008b). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects

models. *Frontiers in psychology*, *4*.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (p. 901-938). Amsterdam: Elsevier.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441-455.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-39)

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201-211.

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer [computer program]*. Retrieved from `http://www.praat.org/` (Version 5.3.18)

Borden, G. J., Raphael, L. J., & Harris, K. S. (1994). *Speech science primer: Physiology, acoustics, and perception of speech* (3rd ed.). Baltimore, MD: Lippincott, Williams & Wilkins.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*(2), 123-147.

Bradlow, A., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707-729.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*(2), 274-296.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383-398.

Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Acoustic determinants of perceptions of personality from speech. *International Journal of the Sociology of Language*, *1975*(6), 11-32.

Brunelliére, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, *125*(1), 82-93.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (Eds.), *Proceedings of the speech prosody 2002 conference* (p. 199-202). Aix-en-Provence: Laboratoire Parole et Langage.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177-208.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535-544.

Christenfeld, N. (1996). Effects of a metronome on the filled pauses of fluent speakers. *Journal of Speech, Language and Hearing Research*, *39*(6), 1232-1238.

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73-111.

Clarke, C., & Garrett, M. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*, 3647-3658.

Collard, P. (2009). *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech*. Unpublished doctoral dissertation, The University of Edinburgh.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *34*(3), 696-702.
Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS One*, *6*(5), e19792.
Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658-668.
Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989-999.
Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862-2873.
Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292-314.
Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon: Multilingual Matters.
De Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, *13*(1), 1-24.
De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *Proceedings of the 6th workshop on disfluency in spontaneous speech (DiSS)* (p. 17-20). Stockholm.
De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). L2 fluency: speaking style or proficiency? Correcting measures of L2 fluency for L1 behavior. *Applied Psycholinguistics*.
De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012a). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5-34.
De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012b). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*.
DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.
Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655-679.
Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S. Wang (Eds.), *Individual differences in language ability and language behavior* (p. 85-101). New York: Academic Press.
Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.
Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709-738.
Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, *29*(2), 320-326.
Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161-175.
Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (p. 123-148). Amsterdam: John Benjamins Publishing.
Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (p. 243-265). Michigan: The University of Michigan Press.
Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral

production. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 215-240.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379-399.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*(5), 501-518.

Goldman-Eisler, F. (1958a). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*(3), 226-231.

Goldman-Eisler, F. (1958b). Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology*, *10*(2), 96-106.

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787-814.

Hanulíková, A., Van Alphen, P., Van Goch, M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878-887.

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, *57*(3), 169-177.

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with 'articulatory' pauses. *Language and Speech*, *26*(3), 203-214.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence and the foreign language teacher.* (p. 243-265). Skokie, Illinois, USA: National Textbook Company.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151-171.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229-249.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. F. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common european framework of reference for languages (CEFR). *Language Testing*, *29*(2), 202-220.

IELTS. ([n.d.]). *IELTS Speaking band descriptors.* Retrieved from `http://www.ielts.org/PDF/UOBDs_SpeakingFinal.pdf` (November 2013)

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(4), 824-843.

Kahng, J. (2013). *Investigating utterance fluency and cognitive fluency in second language speech.* Presentation at the New Sounds 2013 conference. Montréal, Canada.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral english. *The Modern Language Journal*, *94*(4), 554-566.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch

word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643-650.
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38.
Kidd, C., White, K. S., & Aslin, R. N. (2011a). Learning the meaning of "um". toddlers' developing use of speech disfluencies as cues to speakers' referential intentions. In I. Arnon & E. Clark (Eds.), *Experience, variation and generalization: Learning a first language (trends in language acquisition research)* (p. 91-106). Amsterdam: John Benjamins Publishing.
Kidd, C., White, K. S., & Aslin, R. N. (2011b). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925-934.
Kircher, T. T. J., Brammer, M. J., Levelt, W. J. M., Bartels, M., & McGuire, P. K. (2004). Pausing for thought: engagement of left temporal cortex during pauses in speech. *NeuroImage*, *21*(1), 84-90.
Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning*, *49*(2), 303-342.
Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.
Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*(6), 618-625.
Krauss, R. M., & Pardo, J. S. (2006). Speaker perception and social behavior: bridging social psychology and speech science. In P. Van Lange (Ed.), *Bridging social psychology: Benefits of transdisciplinary approaches.* (p. 273-278). Hillsdale, NJ: Erlbaum.
Kutas, M., DeLong, K., & Smith, N. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Using our past to generate a future* (p. 190-207). Oxford University Press.
Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, *32*(2), 389-416.
Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387-417.
Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (p. 25-42). Michigan: University of Michigan Press.
Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41-104.
Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-38.
MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, *111*(1), 36-45.
MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Investigating the consequences of disfluent silent pauses in speech for listeners. *Neuropsychologia*, *48*, 3982-3992.
Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*, 19-44.
Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71-102.
Martin, J. G., & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Perception & Psychophysics*, *3*(6), 427-438.
McColl, D., & Fucci, D. (2006). Measurement of speech disfluency through magnitude estimation and interval scaling. *Perceptual & Motor Skills*, *102*(2), 454-460.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76.

Merlo, S., & Mansur, L. (2004). Descriptive discourse: topic familiarity and disfluencies. *Journal of Communication Disorders*, *37*(6), 489-503.

Mora, J. C. (2006). Age effects on oral fluency development. *Age and the rate of foreign language learning*, 65-88.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5), 453-467.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, *48*(2), 159-182.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451-468.

Oomen, C. C., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, *30*(2), 163-184.

Oostdijk, N. (2000). The spoken Dutch corpus project. *ELRA Newsletter*, *5*(2), 4-8.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, *9*(1), 19-35.

Panico, J., Healey, E. C., Brouwer, K., & Susca, M. (2005). Listener perceptions of stuttering across two presentation modes: A quantitative and qualitative approach. *Journal of fluency disorders*, *30*(1), 65-85.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism* (Vol. 18). John Benjamins Publishing.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-189.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105-110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *FirstView*, 1-19.

Pinget, A.-F., Bosker, H. R., Quené, H., & De Jong, N. H. (in press). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* Springer Verlag.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Reviews in Neuroscience*, *13*, 25-42.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1-2), 103-121.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413-425.

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Raupach, M. (1983). Analysis and evaluation of communication strategies. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (p. 199-209). London, United Kingdom: Longman.

Riazantseva, A. (2001). Second language proficiency and pausing. *Studies in Second Language Acquisition*, *23*, 497-526.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423-441.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review/La revue canadienne des langues vivantes*, *65*(3),

395-412.
Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*(3), 362-367.
Schachter, S., Rauscher, F., Christenfeld, N., & Crone, K. T. (1994). The vocabularies of academia. *Psychological Science*, *5*(1), 37-41.
Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227-256.
Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(4), 357-385.
Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the twenty-eighth meeting of the Cognitive Science Society* (p. 750-755).
Segalowitz, N. (2010). *Cognitive bases of second language fluency.* New York: Routledge.
Segalowitz, N., & Hulstijn, J. H. (2005). Automaticity in bilingualism and second language learning. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (p. 371-388). Oxford, UK: Oxford University Press.
Severens, E., Lommel, S., Ratinckx, E., & Hartsuiker, R. (2005). Timed picture naming norms for 590 pictures in dutch. *Acta Psychologica*, *119*(2), 159-187.
Shriberg, E. E. (1996). Disfluencies in switchboard. In *Proceedings of the international conference on spoken language processing, addendum* (p. 11-14).
Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1-14.
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.
Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (p. 207-226). Brussels: University of Brussels Press.
Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications Limited.
Susca, M., & Healey, E. C. (2001). Perceptions of simulated stuttering and fluency. *Journal of Speech, Language and Hearing Research*, *44*(1), 61-72.
Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485-496.
Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.
Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71-79.
Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (p. 239-273). Amsterdam: John Benjamins Publishing.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84-119.
Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, *28*(02), 251-276.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*(3), 443-

467.

Van Berkum, J. J. A., Van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580-591.

Van Turennout, M., Hagoort, P., & Brown, C. M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science*, *280*(5363), 572-574.

Veenker, T. J. G. (2006). *FEP: A tool for designing and running computerized experiments.* (computer software, version 2.4.19)

Veenker, T. J. G. (2012). *The ZEP experiment control application.* Utrecht Institute of Linguistics OTS, Utrecht University, The Netherlands. Retrieved from `http://www.hum.uu.nl/uilots/lab/zep/` (computer software, version 1.2)

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81-94.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*(2), 63-72.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (p. 102-127). Michigan: University of Michigan Press.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272-1288.

Appendices

Appendix A

Appendix A contains supplementary information about the four experiments reported in Chapter 2. Descriptive data of each speech fragment (acoustic characteristics, together with the estimated ratings on fluency, pausing, speed, and repair) are available online: `www.hrbosker.nl/datachapter2`

Literal instructions to participants in the four experiments (in Dutch; English translations given below):

Experiment 1

> *"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op vloeiendheid. Baseer je oordeel telkens op: (1) het gebruik van pauzes: bijv. geen en/of zeer korte stille en gevulde pauzes, of juist zeer veel en/of zeer lange stille en gevulde pauzes; (2) de snelheid van spreken: bijv. zeer langzaam of zeer snel; (3) het gebruik van herhalingen en correcties: bijv. geen of juist zeer veel herhalingen en/of correcties."*
>
> "It is your task to rate the speech fragments on fluency. Base your judgments on: (1) the use of pauses: e.g., none and/or very short silent and filled pauses vs. very many and/or very long silent and filled pauses; (2) the speed of speaking: e.g., very slow vs. very fast; (3) the use of repetitions and corrections: e.g., none vs. very many."

Experiment 2

"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op het gebruik van pauzes. Baseer je oordeel telkens op de hoeveelheid en de lengte van stille en gevulde pauzes."

"It is your task to rate the speech fragments on the use of pauses. Base your judgments on the frequency and the length of silent and filled pauses."

Experiment 3

"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op de snelheid van spreken, bijv. zeer snel of zeer langzaam."

"It is your task to rate the speech fragments on the speed of speaking, e.g., very slow or very fast."

Experiment 4

"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op het gebruik van herhalingen en correcties. Baseer je oordeel telkens op de hoeveelheid herhalingen en correcties."

"It is your task to rate the speech fragments on the use of repetitions and corrections. Base your judgments on the frequency of such repetitions and corrections."

Table 1: Schematical representations of the scales used in Chapter 2.

Experiment 1: fluency		
	What is your judgment of the fluency?	
not fluent at all	* * * * * * * * *	very fluent
Experiment 2: pauses		
	What is your judgment of the use of pauses?	
none and/or very short pauses	* * * * * * * * *	very many and/or very long pauses
Experiment 3: speed		
	What is your judgment of the speech rate?	
not fast	* * * * * * * * *	very slow
Experiment 4: repair		
What is your judgment of the use of repetitions and corrections?		
no repetitions and/or corrections	* * * * * * * * *	very many repetitions and/or corrections

Appendix B

Appendix B contains supplementary information about the two experiments reported in Chapter 3. Descriptive data of each speech fragment (acoustic characteristics, together with the fluency ratings) are available online: `www.hrbosker.nl/datachapter3`

Literal instructions to participants in the two experiments (in Dutch; English translation given below):

> *"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op vloeiendheid. Baseer je oordeel telkens op: (1) het gebruik van pauzes: bijv. geen en/of zeer korte stille en gevulde pauzes, of juist zeer veel en/of zeer lange stille en gevulde pauzes; (2) de snelheid van spreken: bijv. zeer langzaam of zeer snel; (3) het gebruik van herhalingen en correcties: bijv. geen of juist zeer veel herhalingen en/of correcties."*

> "It is your task to rate the speech fragments on fluency. Base your judgments on: (1) the use of pauses: e.g., none and/or very short silent and filled pauses vs. very many and/or very long silent and filled pauses; (2) the speed of speaking: e.g., very slow vs. very fast; (3) the use of repetitions and corrections: e.g., none vs. very many."

Table 2: Schematical representation of the scales used in Chapter 3.

What is your judgment of the fluency?		
not fluent at all	* * * * * * * * *	very fluent

Appendix C

Example recordings of the native speaker and the non-native speaker in Chapter 4 are available online: `www.hrbosker.nl/datachapter4`

Table 3: Items used in all three experiments of Chapter 4, together with frequency and name agreement data.

	Dutch name	English translation	Freq	FreqGroup	NameAgreement
1	accordeon	accordion	1	LowFreq	94
2	neushoorn	rhinoceros	1	LowFreq	90
3	iglo	igloo	1	LowFreq	100
4	eenhoorn	unicorn	1	LowFreq	100
5	stethoscoop	stethoscope	1	LowFreq	93
6	gieter	watering can	1	LowFreq	97
7	ventilator	fan	2	LowFreq	94
8	pompoen	pumpkin	2	LowFreq	94
9	naaimachine	sewing machine	2	LowFreq	95
10	typemachine	typewriter	1	LowFreq	90
11	ananas	pineapple	2	LowFreq	100
12	schommel	swing	2	LowFreq	100
13	sneeuwman	snowman	1	LowFreq	89
14	cactus	cactus	3	LowFreq	100
15	palmboom	palm tree	3	LowFreq	95
16	stofzuiger	vacuum	3	LowFreq	97
17	zaag	saw	3	LowFreq	100
18	trechter	funnel	2	LowFreq	90
19	eekhoorn	squirrel	3	LowFreq	97
20	eskimo	eskimo	3	LowFreq	100
21	tandenborstel	toothbrush	4	LowFreq	95
22	dolfijn	dolphin	4	LowFreq	100
23	krokodil	alligator	5	LowFreq	100
24	puzzel	puzzle	4	LowFreq	98
25	weegschaal	scale	5	LowFreq	100
26	aardbei	strawberry	5	LowFreq	100
27	vleermuis	bat	6	LowFreq	100
28	slak	snail	5	LowFreq	95
29	vlieger	kite	6	LowFreq	100
30	kruiwagen	wheelbarrow	5	LowFreq	97

Continued on following page.

	Dutch name	English translation	Freq	FreqGroup	NameAgreement
31	brug	bridge	52	HighFreq	97
32	regen	rain	55	HighFreq	89
33	bus	bus	58	HighFreq	100
34	ster	star	61	HighFreq	100
35	maan	moon	65	HighFreq	95
36	schoen	shoe	68	HighFreq	100
37	kom	bowl	70	HighFreq	95
38	vis	fish	73	HighFreq	97
39	sigaret	cigarette	74	HighFreq	97
40	baby	baby	79	HighFreq	92
41	trein	train	81	HighFreq	100
42	telefoon	telephone	84	HighFreq	100
43	tand	molar	89	HighFreq	87
44	bloem	flower	94	HighFreq	100
45	koning	king	100	HighFreq	97
46	neus	nose	101	HighFreq	97
47	fles	bottle	112	HighFreq	100
48	bank	bench	114	HighFreq	92
49	trap	stairs	116	HighFreq	97
50	boom	tree	137	HighFreq	100
51	muur	wall	147	HighFreq	94
52	stoel	chair	151	HighFreq	100
53	hond	dog	168	HighFreq	100
54	kerk	church	205	HighFreq	97
55	auto	car	208	HighFreq	100
56	voet	foot	225	HighFreq	100
57	tafel	table	247	HighFreq	100
58	arm	arm	266	HighFreq	95
59	deur	door	376	HighFreq	100
60	hand	hand	1028	HighFreq	97

Appendix D

Example recordings of the native speaker and the non-native speaker in Chapter 5 are available online: `www.hrbosker.nl/datachapter5`

Below, the experimental items of Chapter 5 are listed. These Dutch items were modeled on the English items used in Collard (2009). In each item the different substitution words (e.g., "wond / verwonding / zakdoek") should be interpreted as: NoChange condition / CloseChange condition / DistantChange condition.

1. De dokter keek hoe lang hij nog moest werken. Hij zag dat de patiënt met de wond / verwonding / zakdoek als enige nog in de wachtkamer zat. Een vriendelijke maar strikte verpleegster bracht de jongen de spreekkamer binnen.

2. We vroegen ons allemaal af waar de nieuwe werknemer naartoe ging. Het was duidelijk dat de vrouw met de papieren / documenten / aktetas een klein beetje verdwaald was. In een groot complex is het gemakkelijk de weg kwijt te raken.

3. Teun was helemaal op de hoogte van de beroemdheden bij de Oscar ceremonie. Blijkbaar werd de film over de aliens / marsmannetjes / dinosauriërs universeel geprezen. Iedereen had het een geweldige ceremonie gevonden.

4. Simon moest echt een beslissing gaan maken over zijn carrière. Hij zei dat de baan uit het tijdschrift / blad / nieuwsbericht interessant had geleken. Hij zocht een baan in de financiële sector.

5. De politie wist nog steeds niet wat ze met het onderzoek aan moesten. Ze dachten dat de jongen met de aansteker / lucifer / knuppel een waarschijnlijke verdachte was. De getuigen hadden geen bruikbare informatie opgeleverd.

6. Uiteindelijk kwamen we erachter wat de nieuwe buurman gedaan had. De boom die de straat / steeg / zon had geblokkeerd, was neergehaald. Het zou echt een enorm verschil voor zijn kleine tuin kunnen zijn.

7. De nieuwe journalist wist niet helemaal zeker wat hij nu moest doen. Hij wist dat het artikel over de inbraak / diefstal / ontvoering allang over de deadline was. Maar de hoofdredacteur had het echt nodig voor de voorpagina.

8. De advocaat vroeg zich af hoe hij deze zaak goed moest verdedigen. Kennelijk was het dossier over het gebouw / pand / bestuur van cruciale waarde. Hij kon het zich niet veroorloven zijn nieuwe cliënt teleur te stellen.

9. De taxichauffeur wist niet precies waar hij zich bevond in de straat. Op een of andere manier kwamen de oude appartementen naast de winkel / zaak / vrachtwagen hem bekend voor. Als hij niet opschoot, zou hij de klant verliezen.

10. De jonge secretaresse ging snel na wat nog op het lijstje stond. In ieder geval lag de brief aan de klant / cliënt / burgermeester al op haar baas zijn bureau. Alle kantoorklusjes moesten voor twaalf uur gedaan zijn.

11. De recensent stond volledig achter zijn laatste recensie in een bekend dagblad. Hij dacht dat het toneelstuk over de twee detectives / rechercheurs / piloten zeker een jaar zou draaien. Hij kende het toneelwereldje van binnen en van buiten en had meestal gelijk.

12. De luchtverkeersleider ging zorgvuldig na of de vlucht volgens plan verliep. Het vliegtuig dat de lading / vracht / koningin zou vervoeren, naderde het vliegveld al. Het kon best wel eens een spannend klusje worden.

13. Een betrouwbare reclame-specialist legde uit hoe men het beste de doelgroep kon bereiken. Hij zei dat de advertentie met de kat / poes / aap gegarandeerd succesvol zou zijn. Hij had een heleboel ervaring binnen de reclame-sector.

14. De wandelaars dachten dat ze de eindbestemming al naderden. Het leek alsof het pad langs de gracht / vaart / weide in de juiste richting liep. Maar zonder een goede kaart was het onmogelijk hier zeker van te zijn.

15. Het werd duidelijk hoe de houding van burgers in de grote stad begon te veranderen. Het nieuws over de oorlog / strijd / staking in het zuiden van het land had iedereen beangstigd. Hoe dan ook zou het leger in de volgende maanden een belangrijke rol gaan spelen.

16. De brandweerman vroeg ons hoe het incident was begonnen. We wezen hem op de vrouw met de kinderwagen / buggy / sjaal die alarm had geslagen. Hij wilde het hele verhaal nog eens van haar horen.

17. Het arrestatieteam ging ervan uit dat de gevaarlijke crimineel nog niet ver weg kon zijn gevlucht. Binnen vijf minuten was het hele gebied rondom de kathedraal / kerk / universiteit afgezet. Desondanks werd hij niet gevonden en een moeizame zoektocht is nog steeds gaande.

18. Eerst snapte ik niet waarom Lea precies deze krappe zitplekken had uitgekozen. Ik dacht dat de stoelen bij de ingang / uitgang / paal beter zicht op het podium hadden. Uiteindelijk bleek het zicht en het toneelstuk prima te zijn.

19. Het leek wel of het hele dorp was gekomen. Alle aanwezigen bij de begrafenis / uitvaart / bijeenkomst waren erg aangedaan door het recente sterfgeval. De dominee schatte in dat er zeker tweehonderd gasten waren geweest.

20. De dierenarts kwam in de wachtruimte kijken wat al dat lawaai betekende. Iemand had de kat met de verwonde poot / klauw / staart opgepakt. De man schreeuwde het uit toen de wilde kat hem in zijn armen krabde.

21. De studente vroeg een oude professor om advies over haar vakken. Hij vertelde dat het boek over Middeleeuwse rituelen / gebruiken / veldslagen essentieel was. Op basis van zijn advies was ze overtuigd dat ze het tentamen kon halen.

22. Het hoofd van het museum wilde op de hoogte worden gehouden van de verhuizing. Het bleek dat een grote kist met het wereldberoemde portret / schilderij / beeld nog steeds in de vrachtwagen lag. Als er iets kwijt zou raken, zou hij zeker razend worden.

23. De student moest nu echt een weloverwogen keuze maken. Een strenge coördinator vertelde dat het vak over scheikunde / natuurkunde / taalkunde was geannuleerd. Hierdoor kwam hij net een aantal studiepunten tekort.

24. De dierenverzorger wist dat hij de hele dag bezig zou zijn. Al dagen lang klaagden bezoekers dat het hok van de panters / jaguars / adelaars zo stonk. Het was een flinke klus, maar gelukkig had hij de hulp van zijn aardige collega's.

25. Uiteindelijk kwam ze erachter waarom hij in grote paniek was. Het bleek dat iemand 's nachts het raam van zijn woning / flat / wagen had ingeslagen. Het incident zou zeker met alle bewoners besproken moeten worden.

26. Iedereen in de enorme bibliotheek vroeg zich af waarom hij zo laat was. Uiteindelijk bleek dat de tas van de auteur / schrijver / arts uitgebreid was doorzocht. De bewaking is altijd extra waakzaam bij dit soort belangrijke bijeenkomsten.

27. Het meisje hoopte dat ze vandaag niet al te veel huiswerk zou hebben. Een volle tas met haar schoolspullen had ze op de vloer / grond / kledingkast gelegd. Ze had altijd een hekel aan het vak wiskunde.

28. De kleine matroos was blij om weer aan land te zijn. Het mankement aan zijn schip / vaartuig / dek moest grondig gerepareerd worden. Hij was van plan om veel familie en oude vrienden te bezoeken.

29. Overal had het meisje al gezocht naar de dure tickets. Ze was ervan overtuigd dat ze kaartjes voor de uitvoering / voorstelling / lezing op de kast had gelegd. Als ze ze niet snel zou vinden, zou ze echt veel te laat zijn.

30. De redactrice slaakte een diepe zucht van opluchting toen ze de oprit opreed. Het vakantiehuis aan het meer / water / strand voelde altijd als thuis. De laatste tijd was ze zo druk geweest dat ze uitkeek naar dit weekendje weg.

31. De brouwer was altijd nieuwe brouwsels aan het uitproberen. Dit keer was het vat met het pils / bier / mengsel daadwerkelijk gaan gisten. Misschien kon hij deze nieuwe uitvinding wel op de markt gaan brengen.

32. De brandweermannen zochten overal naar overlevenden tussen het puin. De oude hut midden in het woud / bos / gebergte was al jaren verlaten. Er was haast niets over gebleven van het kleine houten huisje.

33. Dit jaar was de boer veel beter voorbereid op mogelijke tegenslagen. Hij had zijn bedrijf al klaargemaakt voor een eventuele watervloed / overstroming / storm later in het jaar. Een goede oogst was cruciaal voor het voortbestaan van zijn familiebedrijf.

34. Vorig jaar had het museum nog een nieuw alarmsysteem laten plaatsen. De voetafdrukken op het gazon / grasveld / dak lieten precies zien hoe de sluwe dief naar binnen was gekomen. De dure kunstwerken waren wel verzekerd, maar desondanks onvervangbaar.

35. De jeugdige atleet had veel moeite zijn emoties te bedwingen. De menigte die zich bij het sportterrein / sportveld / vliegveld had verzameld, ging uit zijn dak. Hij was ontzettend gespannen, maar toch genoot hij van de aandacht.

36. De leiders van beide landen ontmoetten elkaar op de geheime plek. Over een belangrijke voorwaarde van het pact / contract / optreden moest nog steeds onderhandeld worden. Het leek erop alsof beide partijen doordrongen waren van de belangen.

Samenvatting in het Nederlands

Spreken is *uh..* zilver

Op 17 april 2013 werd toenmalig kroonprins Willem-Alexander, thans koning, geïnterviewd over de op handen zijnde troonswisseling. Op een gegeven moment stelde een van de interviewers een lastige vraag, waarop de prins antwoordde:

> *"Nee, dit lijkt me echt iets wat niet verstandig is om hier een antwoord op te geven. Ik heb ook wel vaker in interviews gezegd: spreken is zilver, zwijgen is goud."*

Tenminste, dat was zijn antwoord als we het transcript van het interview moeten geloven. Als men echter dit specifieke deel van het interview terugluistert, dan klonk zijn daadwerkelijke antwoord ongeveer als volgt:

> *"Nee* ***[uh]*** *dit lijkt me echt .. iets .. wat* ***[uh]*** *niet* ***[uh]*** *verstandig is om* ***[uh]*** *hier een* ***[uh]****,* ***een, een, een, een*** *antwoord op te geven.* ***[Uh...]*** *Ik heb ook wel vaker in interviews gezegd* ***[uh]****: 'spreken is zilver, zwijgen is goud'.* ***[Uh...]****"*
>
> → http://www.youtube.com/watch?v=DsX4nhOwGBU

Deze gesproken uiting is een duidelijk voorbeeld van een gebrek aan vloeiendheid. Maar niet alleen koninklijke sprekers hebben moeite met het produceren van vloeiende spraak: iedereen heeft wel eens de *uhm*'s geteld van een saaie leraar, of zich geërgerd aan een haperende nieuwslezer(es). Spontane spraak bevat allerlei soorten zogenaamde 'haperingen', zoals stille pauzes, gevulde pauzes (*uh*'s en *uhm*'s), correcties, herhalingen (*"een, een, een, een"*), enz. Toegegeven, het hierboven weergegeven citaat is een vrij extreem voorbeeld van niet-vloeiende spraak (bijna de helft van de totale duur van het antwoord

bestaat uit *uhm*'s en herhalingen). Niettemin schat men dat in spontane spraak er ongeveer zes haperingen per honderd woorden voorkomen (Fox Tree, 1995).

Maar als spontane gesprekken inderdaad zoveel haperingen bevatten, welke invloed heeft dat niet-vloeiende karakter van spraak dan op het begrip bij luisteraars? De wetenschappelijke literatuur lijkt een schijnbaar tegenstrijdig antwoord op deze vraag te geven. Aan de ene kant zijn er studies die stellen dat haperingen in spraak een negatief effect hebben op de evaluatie van vloeiendheid. Met andere woorden: hoe meer haperingen, hoe lager het vloeiendheidsoordeel. Deze studies behoren tot een groep die we in deze dissertatie aanduiden met de *evaluatieve benadering* van vloeiendheid. Binnen deze benadering wordt vloeiendheid geïnterpreteerd als een component van de algehele spreekvaardigheid van de spreker. Deze benadering houdt zich vrijwel uitsluitend bezig met de evaluatie van de vloeiendheid van spraak van tweedetaalsprekers (T2-sprekers). Onderzoekers pogen hierin een valide en betrouwbare manier te vinden om de algehele spreekvaardigheid van een T2-spreker te meten.

Aan de andere kant zijn er ook studies die suggereren dat haperingen positieve effecten kunnen hebben op het begrijpen van spraak. Deze studies kenmerken zich door een benadering die we in deze dissertatie de *cognitieve benadering* van vloeiendheid noemen. Het doel van de cognitieve benadering van vloeiendheid is vast te stellen welke cognitieve factoren verantwoordelijk zijn voor haperingen in spraak (productie), en te begrijpen hoe deze haperingen cognitieve processen in het spraakbegrip van luisteraars beinvloeden (perceptie), zoals aandachtsmechanismen, geheugen, en voorspelling. Zo is er bijvoorbeeld aangetoond dat, als een gesproken uiting een *uhm* bevat, luisteraars (i) de inhoud van de uiting beter onthouden; (ii) sneller reageren op instructies; en (iii) specifieke verwachtingen hebben met betrekking tot het woord dat volgt op de *uhm*. In al deze studies werd gewerkt met moedertaalsprekers (T1-sprekers).

Onderzoeksvraag

Samenvattend kunnen we stellen dat de evaluatieve benadering negatieve effecten van T2-haperingen op de perceptie van vloeiendheid vindt, terwijl de cognitieve benadering positieve effecten van T1-haperingen op spraakbegrip rapporteert. De studies in deze dissertatie trachten deze schijnbare tegenstelling op te helderen door te onderzoeken welk effect vloeiendheidskenmerken in zowel T1- en T2-spraak hebben (i) op de subjectieve vloeiendsheidoordelen van luisteraars, en (ii) op de cognitieve processen die een rol spelen bij het begrijpen van spraak, zoals voorspelling, geheugen, en aandacht. De volgende hoofdonderzoeksvraag wordt geformuleerd:

> Hoofdonderzoeksvraag: Hoe beïnvloeden vloeiendheidskenmerken de perceptie van T1- en T2-spraak?

Hoofdstuk 2 en 3 bestuderen de perceptie van vloeiendheid in T1- en T2-spraak vanuit de evaluatieve benadering. De experimenten in deze hoofdstukken onderzoeken het effect van specifieke soorten haperingen op de *evaluatie* van vloeiendheidskenmerken in T1- en T2-spraak (d.w.z., het subjectieve vloeiendheidsoordeel van luisteraars). Hoofdstuk 4 en 5 bestuderen de perceptie van vloeiendheid in T1- en T2-spraak vanuit de cognitieve benadering. De experimenten in deze hoofdstukken onderzoeken het effect van haperingen zoals *uhm*'s op de *verwerking* van vloeiendheidskenmerken in T1- en T2-spraak (te weten het effect van *uhm*'s op luisteraars' verwachtingen en aandacht).

Wat maakt spraak vloeiend?

Bevindingen van Hoofdstuk 2

Hoofdstuk 2 onderzoekt hoe luisteraars het vloeiendheidsniveau van T2-sprekers beoordelen. We wilden bepalen welke akoestische dimensie de grootste rol speelt bij het bepalen van een subjectief vloeiendheidsoordeel. Letten luisteraars het meest op het gebruik van pauzes, spreeksnelheid, of op herhalingen en correcties? Daarom luidde de eerste onderzoeksvraag:

> Onderzoeksvraag 1A: Welke rol spelen drie afzonderlijke akoestische dimensies (pauzes, spreeksnelheid, en herhalingen en correcties) bij het beoordelen van vloeiendheid?

Voor alle vier de experimenten maakten we gebruik van steeds dezelfde set T2-spraakopnames, verkregen uit het 'What Is Speaking Proficiency'-project (WISP) van de Universiteit van Amsterdam (zoals beschreven in De Jong et al., 2012a). In Experiment 1 werden deze T2-spraakopnames voorgelegd aan een groep luisteraars die de taak hadden de T2-spraak te beoordelen op vloeiendheid. Hiervoor ontvingen de luisteraars specifieke beoordelingsinstructies (zie Appendix A). Deze subjectieve oordelen werden gerelateerd aan objectieve akoestische maten van de T2-spraakmaterialen. Deze akoestische maten werden geclusterd in drie verschillende dimensies (zie Tabel 2.2): *pauzes* (het aantal gevulde pauzes, het aantal stille pauzes, en de gemiddelde duur van de stille pauzes), *spreeksnelheid* (de gemiddelde duur van een lettergreep), en *herhalingen en correcties* (het aantal correcties, en het aantal herhalingen). Deze specifieke akoestische maten werden gekozen voor hun lage intercollineariteit: kruiscorrelaties toonden aan dat de akoestische maten zowel binnen de dimensies als tussen de verschillende dimensies grotendeels onafhankelijk waren (zie Tabel 2.3). Hierdoor kon de afzonderlijke bijdrage aan subjectieve vloeiendheidsoordelen van elk van de verschillende akoestische dimensies met elkaar vergeleken worden.

De resultaten van Experiment 1 tonen ten eerste aan dat de subjectieve oordelen grotendeels waren gebaseerd op de akoestische karakteristieken van de T2-spraak: 84% van de variantie van de subjectieve vloeiendheidsoordelen kon verklaard worden op basis van de totale set van zes akoestische maten. Ten tweede vonden we dat de pauzematen de grootste bijdrage leverden aan vloeiendheidsperceptie (59% verklaarde variantie), gevolgd door spreeksnelheid (54%). De herhalingen en correcties in het T2-spraaksignaal waren slechts in mindere mate gerelateerd aan de vloeiendheidsoordelen (16%).

Vervolgens zochten we naar een mogelijke verklaring voor de resultaten van Experiment 1. We hypothetiseerden dat de belangrijke bijdragen van pauzes en spreeksnelheid aan het beoordelen van vloeiendheid te wijten zouden kunnen zijn aan een mogelijk verhoogde sensitiviteit voor het waarnemen van pauzes en snelheid (tegenover herhalingen en correcties). Bijvoorbeeld, als luisteraars zeer gespitst zijn op het opmerken van pauzes in een spraaksignaal, zou dat de grote rol van pauzes bij vloeiendheidsperceptie kunnen verklaren. Middels drie nieuwe experimenten trachtten we een antwoord te vinden op de tweede onderzoeksvraag:

> Onderzoeksvraag 1B: In welke mate kunnen luisteraars de pauzes, de spreeksnelheid, en de herhalingen en correcties in spraak beoordelen?

De drie nieuwe experimenten maakten gebruik van dezelfde T2-spraakmaterialen als Experiment 1. Echter, de deelnemers kregen verschillende instructies: men werd ofwel gevraagd het gebruik van pauzes te beoordelen (Experiment 2), ofwel de spreeksnelheid (Experiment 3), ofwel het gebruik van herhalingen en correcties (Experiment 4). Vervolgens werden de subjectieve oordelen gerelateerd aan objectief gemeten akoestische kenmerken van de T2-spraak. De hoogste verklaarde variantie werd gevonden voor de pauzeoordelen (Experiment 2): 70% van de oordelen op het gebruik van pauzes kon verklaard worden door de daadwerkelijk gemeten pauzes. Daaruit bleek dat luisteraars het accuraatst zijn als ze spraak beoordelen op het gebruik van pauzes.

Echter, uit de resultaten van Experiment 3 en 4 bleek dat luisteraars ongeveer even accuraat zijn als ze spraak beoordelen op spreeksnelheid als wanneer ze spraak beoordelen op het gebruik van herhalingen en correcties (53% en 55% verklaarde variantie, respectievelijk). Deze bevinding verklaart niet waarom we in Experiment 1 vonden dat herhalingen en correcties slechts in geringe mate bijdragen aan vloeiendheidsoordelen. Samengenomen lijken de experimenten van Hoofdstuk 2 erop te wijzen dat, ondanks de gevoeligheid die luisteraars blijken te hebben voor herhalingen en correcties, zij hun vloeiendheidsoordelen niet baseren op deze spraakkenmerken. Blijkbaar is er geen directe link tussen luisteraars' sensitiviteit en hun vloeienheidsperceptie. Dit suggereert dat luisteraars, volgend op de waarneming van verschillende akoestische dimensies in

T2-spraak, een afweging maken in welke mate ze deze spraakdimensies bij laten dragen aan hun subjectieve vloeiendheidsoordelen.

De perceptie van T1- en T2-vloeiendheid

Bevindingen van Hoofdstuk 3

Hoofdstuk 3 doet verslag van een tweetal experimenten waarin preciezer onderzocht wordt hoe akoestische vloeiendheidskenmerken gewogen worden bij het bepalen van een vloeiendheidsoordeel door de perceptie van T1- en T2-spraak te vergelijken. Het merendeel van de literatuur over vloeienheidsevaluatie heeft T2-spraak onderzocht; vermoedelijk omdat men aanneemt dat T1-spraak sowieso als vloeiend wordt beoordeeld. Echter, de psycholinguïstische literatuur rapporteert dat er aanzienlijke variatie bestaat tussen T1-sprekers in hun productie van haperingen. Dit roept de vraag op:

> Onderzoeksvraag 2: Evalueren luisteraars de vloeiendheidskenmerken in T1- en T2-spraak op dezelfde manier?

Omdat T1- en T2-spraak op een groot aantal linguïstische aspecten sterk van elkaar verschillen, zijn correlationele analyses ongeschikt voor het vergelijken van de perceptie van T1- and T2-vloeiendheid. Daarom hebben we fonetische manipulaties toegepast op T1- en T2-spraakmaterialen die overeenkwamen op één specifieke akoestische eigenschap (bijv. het aantal pauzes). Hierdoor konden we de bijdrage van deze specifieke akoestische eigenschap aan vloeiendheidsoordelen van T1- en T2-spraak vergelijken. Bovendien heeft deze experimentele methode het bijkomende voordeel dat de afzonderlijke bijdrage van verschillende akoestische factoren bepaald kunnen worden. Zo kan men het effect van één akoestisch kenmerk op vloeiendheidsoordelen (bijv. de duur van pauzes) onderscheiden van het effect van een gerelateerd akoestisch kenmerk (bijv. het aantal pauzes).

In Experiment 1 werd het aantal en de duur van stille pauzes gemanipuleerd. Vooraf was erop toegezien dat de spraakmaterialen overeenkwamen voor T1- en T2-spraak op het aantal stille pauzes per 100 lettergrepen. We creëerden drie manipulatiecondities: NoPauses - verwijdering van alle pauzes van >250 ms; ShortPauses - alle pauzes van >250 ms werden gemanipuleerd zodat zij een duur kregen van tussen de 250-500 ms; en LongPauses - alle pauzes van >250 ms werden gemanipuleerd zodat zij een duur kregen van tussen de 750-1000 ms (zie Tabel 3.2).

Deze gemanipuleerde T1- en T2-spraakmaterialen werden voorgelegd aan een groep luisteraars ter beoordeling van de vloeiendheid. Inspectie van deze subjectieve oordelen toonde aan (1) dat de luisteraars T1-spraakmaterialen vloeiender vonden dan T2-spraakmaterialen; (2) dat zowel een toename in het

aantal pauzes, als een toename in de pauzeduur, een negatief effect had op de waargenomen vloeiendheid; en (3) dat deze effecten van de pauzemanipulaties vergelijkbaar waren tussen de T1- en T2-spraak (zie Figuur 3.1).

Het ontwerp van Experiment 2 leek sterk op Experiment 1, maar in Experiment 2 werden er manipulaties toegepast op de snelheid van spreken - in plaats van pauzemanipulaties. Deze snelheidsmanipulaties werden kruisgewijs toegepast: T2-spraak werd versneld tot de gemiddelde T1-snelheid, en T1-spraak werd vertraagd tot de gemiddelde T2-snelheid. Op deze manier konden de vloeiendheidsoordelen op de gemanipuleerde T1- en T2-spraak vergeleken worden. De resultaten van Experiment 2 leken erg op die van Experiment 1 (zie Figuur 3.2). Wederom werd de T1-spraak over het algemeen als vloeiender beoordeeld dan de T2-spraak. Daarnaast had de vertraging van T1-spraak een negatief effect, en de versnelling van T2-spraak een positief effect op de vloeiendheidsoordelen. Deze effecten van onze snelheidsmanipulaties waren van dezelfde orde van grootte. Op basis van de resultaten van Experiment 1 en 2 concluderen wij dat er geen verschil is in de manier waarop de vloeiendheidskenmerken van T1- en T2-spraak gewogen worden. Daarom is er geen reden om aan te nemen dat luisteraars een kwalitatief verschil maken tussen T1- en T2-spraak wanneer ze de vloeiendheid van verschillende sprekers beoordelen.

Uhm en luisteraars' verwachtingen

Bevindingen van Hoofdstuk 4

Waar Hoofdstuk 2 en 3 de evaluatie van vloeiendheid bestudeerden, onderzochten Hoofdstuk 4 en 5 welke effecten bepaalde haperingen hebben op spraakbegrip. De psycholinguïstische literatuur stelt, bijvoorbeeld, dat haperingen in T1-spraak luisteraars kunnen helpen bij spraakbegrip. Luisteraars kunnen gebruikmaken van bepaalde patronen die aanwezig zijn in de productie van haperingen om zo specifieke linguïstische inhoud te voorspellen. Zo wijst de literatuur over spraakproductie erop dat sprekers vaak haperen voorafgaand aan onverwachte of weinig voorkomende woorden. Daardoor kunnen luisteraars haperingen gebruiken als indicaties van onverwachte linguïstische inhoud.

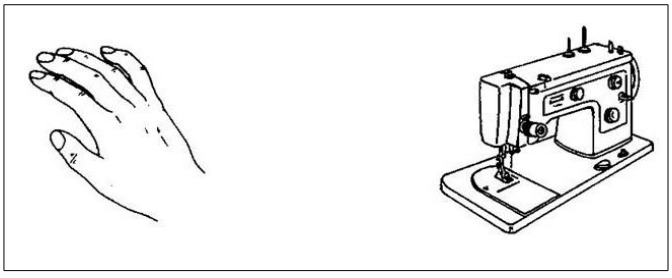
De zojuist genoemde literatuur laat zien hoe luisteraars gebruik kunnen maken van haperingen in T1-spraak. Het is echter niet duidelijk of haperingen in T2-spraak luisteraars ook kunnen helpen bij spraakbegrip. Daarom vergeleek Hoofdstuk 4 de manier waarop T1- en T2-haperingen luisteraars' verwachtingen kunnen sturen. Omdat T2-sprekers op een minder regelmatige manier haperingen produceren in hun T2-spraak, verwachtten we dat haperingen in T2-spraak geen effect hebben op luisteraars' verwachtingen van het volgende woord.

Experiment 1 trachtte de volgende onderzoeksvraag te beantwoorden:

> Onderzoeksvraag 3A: Verwachten luisteraars, bij het horen van een hapering, dat de spreker een laag-frequent woord zal gaan uitspreken?

Hiervoor maakten we gebruik van oogbewegingsregistratie. De methode was als volgt (overgenomen en aangepast van Arnold et al., 2007): plaatjes van hoog-frequente (bijv. een hand) en laag-frequente objecten (bijv. een naaimachine) werden getoond op een computerscherm (zie Figuur 1). Tegelijkertijd ontving de deelnemer vloeiende en niet-vloeiende instructies om op een van de plaatjes te klikken (bijv. 'Klik op de rode [...]' vs. 'Klik op *uh* de rode [...]'). We verwachtten dat luisteraars de aanwezigheid van een hapering (*uh*) zouden toeschrijven aan moeite bij de spreker om het woord voor het laag-frequente object te formuleren. Dit zou dan af te meten moeten zijn aan een groter aantal fixaties op het laag-frequente object, nog voordat het laatste woord (nl. het doelwoord) zou worden uitgesproken.

Figuur 1: Voorbeeld van visuele stimuli zoals gebruikt in Experiment 2-3, bestaande uit een hoog-frequent (hand) en een laag-frequent object (naaimachine).



Experiment 1 slaagde er niet in bewijs te leveren voor onze hypothese: haperingen in T1-spraak hadden geen enkel effect op luisteraars' oogbewegingen. Twee mogelijke factoren werden door ons aangewezen die mogelijk hiervoor verantwoordelijk waren: (1) een familiarisatie-fase voorafgaand aan het oogbewegingsexperiment; en (2) de temporele afstand tussen de hapering *uh* en het doelwoord (een bijvoeglijk naamwoord scheidde de hapering van het doelwoord). Daarom werd er gekozen om het experiment te herhalen, maar ditmaal zonder een familiarisatie-fase en met kortere instructiezinnen (d.w.z., zonder bijvoeglijk naamwoorden: 'Klik op de [...]' vs. 'Klik op *uh* de [...]').

De resultaten van Experiment 2 lieten zien dat, als luisteraars een vloeiende zin hoorden, zij geen voorkeur voor een van beide plaatjes hadden (proportioneel evenveel blikken naar hoog- en laag-frequente objecten). Echter, op het moment dat luisteraars een hapering (*uh*) hoorden, keken zij bij voorkeur naar het laag-frequente plaatje (proportioneel meer blikken naar laag-frequente objecten). Deze resultaten suggereren dat luisteraars de waargenomen haperingen toeschreven aan moeite bij de spreker om laag-frequente woorden te formuleren.

Experiment 3 zette het onderzoek van Experiment 2 voort door de effecten van haperingen in T1-spraak (Experiment 2) te vergelijken met die van haperingen in T2-spraak (Experiment 3):

> Onderzoeksvraag 3B: Hebben haperingen in T1- en T2-spraak hetzelfde effect op luisteraars' verwachtingen?

Experiment 3 was identiek aan Experiment 2, maar ditmaal hoorden deelnemers instructies uitgesproken door een niet-moedertaalspreker van het Nederlands (d.w.z. Nederlands met een sterk buitenlands accent). De resultaten van dit derde experiment toonden geen effect van de haperingen van deze T2-spreker. Met andere woorden, T1-haperingen brachten een verwachting teweeg bij luisteraars dat er een laag-frequent woord zou volgen, maar T2-haperingen hadden niet dit effect. Wij stellen dat de luisteraars in Experiment 3 het gebruik van haperingen reduceerden omdat ze een buitenlands accent hoorden. Hun eerdere ervaringen met haperingen in T2-spraak, en specifiek de onregelmatige patronen waarin deze voorkomen, weerhoudt luisteraars ervan om haperingen in T2-spraak te gebruiken bij het opbouwen van linguïstische verwachtingen. Deze bevinding suggereert dat kennis over de identiteit van de spreker beïnvloedt hoe luisteraars de vorm van een bepaalde uiting (d.w.z. vloeiend of niet-vloeiend) gebruiken bij spraakverwerking.

Uhm en luisteraars' aandacht

Bevindingen van Hoofdstuk 5

Hoofdstuk 5 bestudeerde welke invloed haperingen in T1- en T2-spraak hebben op de aandacht van luisteraars. Eerdere studies hebben beargumenteerd dat gevulde pauzes in T1-spraak (bijv. *uhm*'s) een tijdelijke verhoging van de aandacht van luisteraars tot gevolg kunnen hebben (Collard, 2009; Collard et al., 2008; Corley et al., 2007; MacGregor et al., 2010). Zo vonden deze studies bijvoorbeeld dat woorden die volgden op een *uhm* beter onthouden werden door luisteraars (in vergelijking tot dezelfde woorden maar dan vloeiend uitgesproken). Dit roept de vraag op of haperingen in T2-spraak hetzelfde effect hebben op de aandacht van luisteraars, of niet. Daarom werd de volgende onderzoeksvraag geformuleerd:

> Onderzoeksvraag 4: Hebben haperingen in T1- en T2-spraak hetzelfde effect op de aandacht van luisteraars?

Twee mogelijke antwoorden op deze vraag werden onderscheiden. Aan de ene kant zouden de aandachtseffecten van haperingen veroorzaakt kunnen worden door bepaalde patronen in de productie van haperingen. Haperingen in T1-spraak komen over het algemeen vaak voor voorafgaand aan relatief complexere informatie. Daarom zouden luisteraars profijt kunnen hebben van verhoogde aandacht in reactie op een hapering om zo een efficiënte spraakverwerking te waarborgen. De distributie van haperingen in T2-spraak, daarentegen, is onregelmatiger waardoor deze T2-haperingen slechtere indicatoren zijn van aanstaande complexe informatie. Daarom zou het effect van T2-haperingen op de aandacht van luisteraars minder sterk kunnen zijn dan het effect van T1-haperingen (vgl. de resultaten van Hoofdstuk 4).

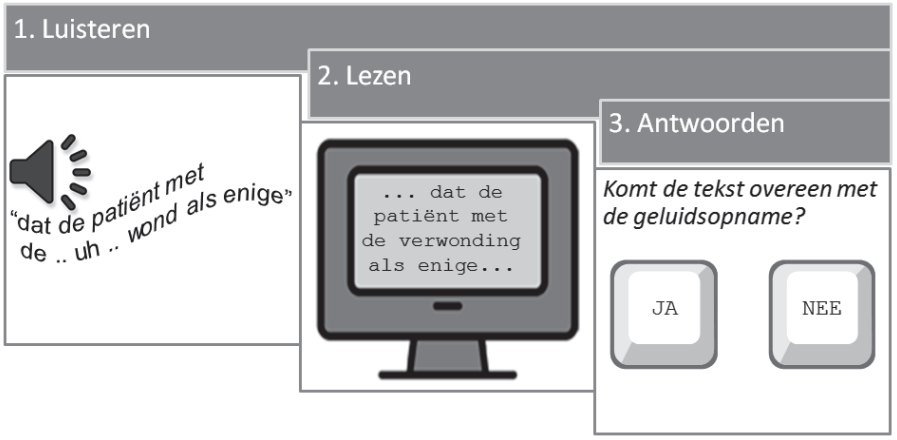
Aan de andere kant zou de verhoogde aandacht bij luisteraars een automatisch cognitief proces kunnen zijn in reactie op oponthoud bij de spreker. De *Temporal Delay Hypothesis* (Corley & Hartsuiker, 2011) stelt dat elke vorm van temporeel oponthoud, waaronder dus ook haperingen, de herkenning en verwerking van woorden bij luisteraars faciliteert. Op basis van deze hypothese zou het effect van T1- en T2-haperingen vergelijkbaar moeten zijn, omdat zowel T1- en T2-haperingen een oponthoud in de spraak vormen.

In Hoofdstuk 5 onderzochten we de effecten van T1- en T2-haperingen op de aandacht van luisteraars door middel van het *Change Detection Paradigm* (CDP; zie ook Figuur 2). Deelnemers werd gevraagd om een gesproken passage van drie zinnen te beluisteren en te onthouden, zoals:

> *"De dokter keek hoe lang hij nog moest werken. Hij zag dat de patiënt met de [uh]* ***wond*** *als enige nog in de wachtkamer zat. De vriendelijke maar strikte verpleegster bracht de jongen de spreekkamer binnen.*

Een van de woorden in de gesproken passage (het zogenaamde doelwoord) werd uitgesproken in een vloeiende context, dan wel in een niet-vloeiende context ("... dat de patiënt met de *wond* ..." vs. "... dat de patiënt met de *uh wond* ..."). Volgend op de gesproken passage, werd op een computerscherm een transcript van de passage gepresenteerd. Echter, soms bevatte dit transcript een substitutie: het doelwoord (wond) was dan in het transcript vervangen door een sterk gelijkend woord (verwonding). De deelnemers hadden de taak om aan te geven of het transcript klopte met de gesproken passage (d.w.z. is er een woord vervangen of niet?). Als haperingen inderdaad de aandacht van luisteraars richten op het daaropvolgende woord, zouden we verwachten dat luisteraars beter zijn in het detecteren van substituties van woorden die in een

Figuur 2: Schematische representatie van het Change Detection Paradigm.



niet-vloeiende context waren uitgesproken ("de *uh wond*"), in vergelijking met woorden die in een vloeiende context voorkwamen ("de *wond*").

Twee experimenten werden ontworpen: Experiment 1 onderzocht de verwerking van haperingen in T1-spraak en Experiment 2 onderzocht de verwerking van haperingen in T2-spraak. De antwoorden van de deelnemers (zie Figuur 5.2) toonden aan dat haperingen een positief effect op de accuraatheid van de deelnemers hadden: vervanging van woorden in een niet-vloeiende context werd beter waargenomen dan vervanging van woorden in een vloeiende context. Dit effect van haperingen werd gevonden in beide experimenten: zowel haperingen in T1-spraak als haperingen in T2-spraak leidden tot een verhoogde aandacht bij luisteraars. Er werd geen bewijs gevonden voor een gereduceerd effect van haperingen in T2-spraak.

Deze bevindingen suggereren dat de identiteit van de spreker geen effect heeft op hoe luisteraars haperingen verwerken. Dit zou in kunnen houden dat luisteraars, bij het horen van een hapering, hun aandacht verhogen op een relatief automatische wijze (vgl. de *Temporal Delay Hypothesis*; Corley & Hartsuiker, 2011). Echter, in Hoofdstuk 5 worden ook enkele bezwaren genoemd met betrekking tot de methodologie (bijv. het feit dat er gebruik werd gemaakt van 'voorgelezen spraak'). Daarom pleiten wij voor enige terughoudendheid in het trekken van conclusies op basis van deze data. Ondanks deze bezwaren onderstrepen de experimenten van Hoofdstuk 5 de belangrijke rol die aandacht speelt bij het verwerken van niet-vloeiende spraak.

Conclusie

Het onderzoek in deze dissertatie werd gemotiveerd door een schijnbare tegenstelling tussen de evaluatieve benadering (negatieve effecten van T2-haperingen op vloeiendheidsoordelen) en de cognitieve benadering van vloeiendheid (positieve effecten van T1-haperingen op spraakbegrip). De studies in deze dissertatie trachten deze schijnbare tegenstelling op te helderen door de consequenties van vloeiendheidskenmerken in T1- en T2-spraak voor spraakperceptie te onderzoeken. In het laatste hoofdstuk van deze dissertatie wordt een integrale beschrijving van de perceptie van vloeiendheid in T1- en T2-spraak voorgesteld.

Deze beschrijving veronderstelt dat een gebrek aan vloeiendheid veroorzaakt wordt door cognitieve inefficiëntie van de processen die betrokken zijn bij spraakproductie. Zowel T1- als T2-sprekers hebben soms moeite met spraakproductie, omdat beiden gebonden zijn aan de tijdsdruk waaronder natuurlijke gesprekken plaatsvinden. Een stille pauze kan voor T1- en voor T2-spraak net zo goed een symptoom zijn van cognitieve inefficiëntie en daarom wegen luisteraars de vloeiendheidskenmerken van T1- en T2-spraak even zwaar (vgl. Hoofdstuk 3).

Echter, dit betekent niet dat haperingen in T1-spraak en T2-spraak ook op dezelfde manier verwerkt worden. Hoofdstuk 4 toonde aan dat luisteraars gebruik kunnen maken van symptomen van cognitieve inefficiëntie omdat deze aanwijzingen geven over de volgende inhoud. Wanneer luisteraars naar een T1-spreker luisteren, beïnvloeden de haperingen in het spraaksignaal hun verwachtingen over de informatie die volgt op een hapering. Wanneer luisteraars daarentegen naar een T2-spreker luisteren, hebben haperingen geen invloed op luisteraars' verwachtingen. Dit suggereert dat de identiteit van de spreker het effect van haperingen op spraakbegrip kan moduleren. Een verklaring voor deze modulatie wordt gevonden in de distributie van haperingen in T1- en T2-spraak. Waar haperingen in T1-spraak bepaalde patronen volgen, is de productie van haperingen in T2-spraak minder regelmatig. Omdat luisteraars ervaring hebben met de onregelmatige verdeling van haperingen in T2-spraak, reduceren zij het effect van die haperingen op hun linguïstische verwachtingen.

Daarnaast hebben T1- en T2-haperingen ook invloed op de aandacht van luisteraars. Een mogelijke verklaring voor dit effect zou gevonden kunnen worden in de *Temporal Delay Hypothesis* (Corley & Hartsuiker, 2011). Deze stelt dat elke vorm van oponthoud in het spraaksignaal de aandacht van luisteraars beïnvloedt. Het oponthoud biedt de luisteraar extra tijd om zich te oriënteren op de aanstaande informatie en zou zo automatisch tot verhoogde aandacht kunnen leiden. Omdat zowel T1- als T2-haperingen oponthoud in het spraaksignaal introduceren, leiden zij beide tot verhoogde aandacht voor de woorden volgend op de hapering. Hoofdstuk 5 kan echter niet onderscheiden tussen ver-

schillende verklaringen voor de gevonden effecten van T1- en T2-haperingen op de aandacht van luisteraars. Nieuwe studies zullen de onderliggende mechanismen verantwoordelijk voor deze aandachtseffecten moeten ophelderen.

De voorgestelde beschrijving van hoe luisteraars omgaan met het gebrek aan vloeiendheid in spontane gesproken communicatie levert belangrijke bijdragen aan verschillende vakgebieden. Allereerst heeft Hoofdstuk 2 een hiërarchie voorgesteld van akoestische dimensies die betrokken zijn bij vloeiendheidsoordelen. Deze hiërarchie is toepasbaar op taaltoetsen. Zo zouden het belang van pauzes en spreeksnelheid (tegenover herstelstrategieën zoals herhalingen en correcties) benadrukt kunnen worden in instructies voor spraakbeoordelaars.

Daarnaast ontkrachten de bevindingen van Hoofdstuk 3 de gangbare aanname dat T1-spraak sowieso als vloeiend wordt waargenomen; veeleer bestaat er beduidende variatie in de vloeiendheid van T1-sprekers. Terwijl veel taaltoetsen T2-sprekers evalueren op grond van een hypothetische T1-norm, betogen wij dat er geen standaard T1-norm bestaat. In plaats daarvan zullen taaltoetsen uiteindelijk een onderscheid moeten kunnen maken tussen die spraakkenmerken die communicatie hinderen en die spraakkenmerken die communicatie bevorderen.

Hoofdstuk 4 en 5 tonen bij uitstek dat vloeiendheidskenmerken in T1- en T2-spraak niet alleen maar communicatie hinderen. In plaats daarvan kunnen luisteraars op een zeer inventieve manier gebruik maken van haperingen in het spraaksignaal. In het bijzonder stellen de resultaten van Hoofdstuk 4 de complexiteit van gesproken communicatie tentoon. Efficiënt begrip van spraak is niet alleen gebonden aan de inhoud van een uiting (zoals de betekenis van de woorden of de grammaticale opbouw van de zin), maar ook aan de vorm van de uiting (bijv. de vloeiendheid). Sterker nog, de cognitieve processen betrokken bij spraakbegrip (zoals verwachtingen opbouwen van aanstaande informatie) worden ook nog eens gemoduleerd door kennis over de identiteit van de spreker.

Deze dissertatie draagt bij aan de studie van gesproken communicatie. De combinatie van de evaluatieve en cognitieve benadering van vloeiendheid vergroot het inzicht in de invloed van vloeiendheidskenmerken op de evaluatie en de verwerking van T1- en T2-spraak. De beschreven resultaten van de huidige studies getuigen van het feit dat de vorm van een spraakuiting een centrale rol speelt bij spraakbegrip: geslaagde communicatie is klaarblijkelijk niet alleen afhankelijk van *wat* er gezegd wordt, maar ook van *hoe* het gezegd wordt en *door wie*.

Acknowledgments

I know it's a cliché and yet it is no less true that this book would not have been written without the support of so many colleagues, friends, and family.

First and foremost, I am indebted to my supervisors: Nivja, Hugo, and Ted.

Nivja, I really couldn't have wished for a better supervisor. From start to finish, you were always available to give advice, answer questions, and provide new motivation. You showed me how to design methodologically sound experiments, and you taught me the value of "check, check, en nog eens check". And I guess I'll never forget our 'work meetings', discussing new experiments over a cappuccino sitting in the sun at Café Van Engelen.

Hugo, thank you for everything you have taught me during my time in Utrecht. If I know anything about statistics and using R, I learned it from you. I have always enjoyed our conversations about GLMMs, LMERs, ANOVAs, and MCMC methods. I very much appreciate the fact that you always made time in your busy schedule to meet up, to comment on a manuscript, or to answer emails - even late at night.

Ted, thank you for your advice during the course of this project. You provided strategic suggestions regarding what study to prioritize, or how to frame a particular manuscript. You also helped in trying to make sense of conflicting or null results. Desalniettemin, I'm most grateful for the Bossche bollen.

I would also like to thank the organizations that made this research possible. Thanks to Pearson Language Testing who funded this project, to the researchers within the 'What Is Speaking Proficiency'-project from Amsterdam who kindly made their speech materials and test scores available, to UiL OTS for providing excellent research facilities and an inspiring academic environment, and to the UiL OTS lab managers. I am particularly grateful to Theo Veenker for his technical support and extraordinary patience, and to Iris Mulders for help with participant recruitment. Thanks also to all (> 500!)

who participated in my silly experiments (voluntarily or under immense social pressure).

Many colleagues made my time at the Trans a time to remember. Thanks to all the people from Taalbeheersing, and to my fellow PhDs in particular: Suzanne and Marloes (never again will you hear another 'Klik op *uh* de blauwe dolfijn'), Anneloes, Louise, Rosie, Ingrid, Naomi, Rogier, Gerdineke, Monica, Renske, Hanna, Björn, Anne, and Marrit. I'm also grateful to many people from UiL OTS for the fun times at conferences abroad, for fruitful discussions at ELiTU, or for just having a nice chat when testing in the catacombs of JKH13: to Brigitta, Sandrien, Rob, Desiree, Hannah, Marijn, and Arnout. Thanks also to Alexis Dimitriadis who made the lotdiss.cls file available for writing this book using LaTeX. Special thanks go to Ileana, Shalom, and Liza, who kindly lend me their voices for some of my experiments, and particularly for their patience during the recordings. And, of course, also many thanks to Heidi, Jade, Farhad, and Indira, for helping out with testing participants.

Ik wil ook een aantal vrienden bedanken voor hun steun tijdens mijn PhD. Zij hielden me met beide voeten op de grond door bijvoorbeeld irritante vragen te stellen over "het algemene nut van je onderzoek?". Tim ("joh, je zou er haast een boek over kunnen schrijven..."), Jan (← Nederlands kampioen schapenscheren), en Manuel ("hoe is het met de *uhm*'s?"): mogen er nog maar vele Black Books-avonden volgen! Benjamin en Benjamin: bedankt dat ik bij jullie altijd met een denkbeeldige PowerPoint-presentatie aan mag komen zetten, en nice cover illustration, by the way! Alle Spiriti: bedankt voor de support. Speciale vermelding verdienen mijn paranimfen: Anne-France (bedankt voor de goeie samenwerking op Hoofdstuk 2, en voor alle koffiemomenten op de Trans) en Bärry (bedankt voor je goeie grappen, je scherpe opmerkingen, ennuh... you're next!).

En natuurlijk ben ik dankbaar voor de steun die ik altijd heb ervaren van mijn (schoon)familie. Pap&mam: geen woorden kunnen uitdrukken hoe dankbaar ik jullie ben voor jullie steun en gebed. Maarten&Maud&Ids&...: bedankt voor de 'ff tussendoor' lunches op de Vismarkt, de Bijleveldstraat, of de cappuccino's bij De Zaak. Annieke: zonder jouw stem - en bovenal je geduld bij allebei (!) de opnamesessies, geen Hoofdstuk 4! Hannebeth: het was altijd goed om bij jou in het hofje te kunnen komen crashen na een werkdag. Ad&Jeannette: altijd leuk om op de Valkenburgerweg over mijn gekke experimentjes te vertellen. Michiel&Rianne: bedankt voor de lijst autowoorden! Justin: bedankt voor het proefdraaien in het lab in Utrecht. Martine: nu ben je niet meer de enige 'auteur' in de familie...

En natuurlijk heel veel dank aan Mirjam. Voor je luisterend oor, voor het meedenken, voor het niet boos worden als ik midden in de nacht een nieuw experiment bedacht had, maar bovenal voor je onvoorwaardelijke liefde.

Curriculum Vitae

Hans Rutger Bosker was born on the 10$^{\text{th}}$ of September, 1987 in Leiderdorp (The Netherlands). In 2004, he obtained his Gymnasium diploma from the Stedelijk Gymnasium in Arnhem. He went on to study Modern Languages at the University of Southampton (United Kingdom), where he received a Certificate of Higher Education. In 2005, he continued his studies in Leiden (The Netherlands), where he obtained his BA and ResMA in Linguistics, specializing in phonetics and psycholinguistics. In 2010, he started his PhD research at the Utrecht institute of Linguistics OTS (UiL OTS). This dissertation is the result of the research he carried out during that period. As of January 2014, he holds a post-doctoral research position at the Max Planck Institute for Psycholinguistics in Nijmegen (The Netherlands).