

Bare Nouns and Beyond
Rethinking Mandarin (In)definiteness
through Translation Corpus Research

Published by

LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
<https://www.lotschool.nl>

Cover illustration: Abstract gradient background by Annie Spratt via Unsplash+

ISBN: 978-94-6093-496-4
DOI: <https://dx.medra.org/10.48273/LOT0711>
NUR: 616

Copyright © 2026: Jianan Liu. All rights reserved.

Bare Nouns and Beyond
Rethinking Mandarin (In)definiteness
through Translation Corpus Research

Voorbij Kale Nomina
Een heroverweging van (on)bepaaldheid
in het Mandarijn
door middel van vertaalcorpusonderzoek
(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. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op
dinsdag 31 maart 2026 des middags te 4.15 uur

door

Jianan Liu

geboren op 23 februari 1990
te Lanzhou, Gansu, China

Promotor:

Prof. dr. H.E. de Swart

Copromotor:

Dr. B.S.W. Le Bruyn

Beoordelingscommissie:

Prof. dr. N.F.M. Corver (voorzitter)

Prof. dr. J.S. Doetjes

Prof. dr. H. de Hoop

Prof. dr. R.P.E. Sybesma

Dr. J. Zwarts

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de China Scholarship Council.

Contents

Acknowledgements	ix
Abbreviations	xi
1 Introduction	1
1.1 From functionalist observations to a formal semantic puzzle	1
1.2 Theoretical background: the Kinds Approach and the Properties Approach	3
1.2.1 The formal analysis of Mandarin argument formation	3
1.2.2 Cross-linguistic perspectives	5
1.2.3 The alternation challenge: summary	7
1.3 Guiding empirical and theoretical questions	8
1.4 Methodology: a translation corpus approach	8
1.5 The argumentation of this dissertation	11
1.6 Conclusion	26
2 A Corpus Investigation of the Mandarin Referential System	29
2.1 Introduction	29
2.2 The translation corpus	31
2.2.1 Corpus construction, data collection, and annotation	31
2.2.2 Benchmark categories	32
2.2.3 Predictions	32
2.3 Results	33
2.3.1 Benchmark categories	33
2.3.2 Overall distribution of referential expressions	37
2.3.3 Testing predictions: alignment in definite and indefinite contexts	39
2.3.4 Recap of observations	46
2.4 Discussion	47
2.5 Conclusion	49

3	“Articleless” Languages are Not Created Equal	53
3.1	Introduction	53
3.2	BNs in “articleless” languages: from Hindi to Russian and Mandarin	54
3.2.1	Chierchia’s predictions for Hindi BSs and BPs	55
3.2.2	Dayal’s account and predictions	56
3.2.3	Towards a cross-linguistic re-assessment of Dayal’s account	58
3.3	Methodology	59
3.4	Results	61
3.4.1	Singular indefinite contexts	61
3.4.2	Singular definite contexts	62
3.4.3	Plural indefinite contexts	63
3.5	Discussion	63
3.5.1	Singular definite contexts	63
3.5.2	Plural indefinite contexts	65
3.5.3	Singular indefinite contexts	67
3.5.4	Recap	72
3.6	Conclusion	73
4	The Theory of Argument Formation: between Kinds and Properties	75
4.1	Introduction	75
4.2	The Kinds Approach and its predictions	78
4.2.1	Non-classifier languages	78
4.2.2	Classifier languages: the case of Mandarin	80
4.2.3	Predictions	80
4.3	The Properties Approach and its predictions	83
4.4	A parallel corpus study	84
4.4.1	Methodology	85
4.4.2	Results	87
4.4.3	Discussion	88
4.4.4	The explanatory potential of the PA and the KA	91
4.4.5	Two notes on methodology	93
4.5	Conclusion	94
5	Fine-Tuning the Property-Based Analysis: the Indefinite Domain	95
5.1	Introduction	95
5.2	On articles and pseudo-incorporation in Mandarin	97
5.3	Diagnosing pseudo-incorporation: Part I	100
5.4	Diagnosing pseudo-incorporation: Part II	103
5.5	Mandarin pseudo-incorporation	106
5.6	Analyzing pseudo-incorporation: capturing typicality and the place of Mandarin pseudo-incorporation	112
5.6.1	Luo (2022)	114
5.6.2	Dayal (2003, 2011)	115
5.6.3	Espinal & McNally (2011)	117
5.6.4	Our analysis	119

5.6.5	Discussion	124
5.7	Conclusion	127
6	Translation Mining: Definiteness across Languages	129
6.1	Introduction	129
6.2	Weak and strong definites across languages: setting the stage	130
6.2.1	Data	130
6.2.2	Analysis	131
6.2.3	Summary and outlook	133
6.3	Corpus and methodology	134
6.3.1	Corpus	134
6.3.2	Methodology	134
6.4	Results and discussion	138
6.4.1	Results	138
6.4.2	Discussion	138
6.5	Two types of strong definiteness	144
6.5.1	Dialectal variation	144
6.5.2	Pragmatic coreference	144
6.5.3	Text-level vs. situation-level familiarity	145
6.5.4	Summary	148
6.6	Conclusion	148
7	Analyzing the Competition between Bare nouns and Demonstratives	151
7.1	Introduction	151
7.2	Anaphoric definites and demonstratives in Ahn (2022)	153
7.2.1	Ahn (2022)	153
7.2.2	A type-shifting implementation of Ahn (2022)	156
7.2.3	Predictions	156
7.3	Introducing our study	158
7.3.1	Corpus	158
7.3.2	Annotation	158
7.3.3	Statistical analysis	161
7.4	Results and discussion	161
7.4.1	Node 1: nominal change	162
7.4.2	Node 3: distance	164
7.4.3	Node 4: uniqueness	165
7.4.4	General discussion	169
7.5	Conclusion	175
8	Conclusion	177
8.1	Thesis Goal	177
8.2	Empirical findings	180
8.3	Theoretical contributions	183
8.4	Methodological innovations	186
8.5	Research limitations and future directions	189

8.5.1	Potential scope of empirical data	189
8.5.2	Methodological triangulation	191
8.5.3	Theoretical exploration	191
8.6	Conclusion	192
A	Distribution of English and Mandarin Referential Expressions in the HP Corpus (dataset for Chapter 2)	195
	Bibliography	201
	Source Text and Translations	211
	Samenvatting in het Nederlands	213
	Curriculum Vitae	219

Acknowledgements

First and foremost, my deepest gratitude goes to my supervisor, Henriëtte de Swart, and my co-supervisor, Bert Le Bruyn. I am sincerely thankful for your invaluable insights and dedicated support throughout this academic journey. To Henriëtte, I still vividly remember my Master's interview, and you were the first person I met from Utrecht. I am deeply grateful for your continued faith in me, from my very first Master's interview to the completion of this PhD. Thank you for opening the doors to Utrecht for me and for your guidance ever since. To Bert, thank you for your immense patience and tolerance for my stubbornness. I am especially grateful for your kindness when I felt lost or raised the simplest of doubts; you never failed to guide me through my confusion with grace. Thank you for encouraging me to live with uncertainty and to keep faith that success lies in persevering through failures without losing enthusiasm. Without the guidance provided by both of you, this dissertation would not have been possible. I cannot thank you enough for your constant support.

I would like to express my sincere thanks to the members of my reading committee for their time and expertise in reviewing this dissertation: Norbert Corver, Jenny Doetjes, Helen de Hoop, Rint Sybesma, and Joost Zwarts. I am truly honored by your willingness to evaluate this work, and I thank you for your insightful comments and for being part of this important milestone.

I would also like to offer special thanks to the HHRM group: Olga Borik, Ljudmila Geist, Daria Seres, Hagay Schurr, Sadhwi Srinivas, and Shravani Patil. Our online and offline meetings in Berlin, Prague, and New Haven were instrumental in shaping this research. It has been a true pleasure and privilege to collaborate with you; your insights and discussions significantly improved the quality of this work.

My gratitude also extends to the members of the *Time in Translation* project: Mar-

tijn van der Klis, Martin Fuchs, Anja Goldschmidt, Chou Mo, and Jos Tellings. Your expertise and dedication during the early stages of my PhD research were a constant source of inspiration.

Life in the Netherlands was made much warmer by the wonderful colleagues and friends I encountered. I would like to thank the following people: Corentin Bourdeau, Quy Ngoc Thi Doàn, Xiaoli Dong, Sybolt Friso, Rachida Ganga, Mengru Han, Na Hu, Shuangshuang Hu, Bambang Kartono, Xin Li, Yuchen Li, Yachan Liang, Ying Liu, Chou Mo (thanks again!), Giada Palmieri, Tijn Schmitz, Joanna Wall, and Yuan Xie. Thank you so much, Maaïke Schoorlemmer, for your constant support throughout my PhD studies in all aspects.

To Sonya Nikiforova: Thank you for being the best office mate who shares the same love for *The Office*.

To Kexin and Jan: Thank you for your caring, heartfelt hugs, and all the sweets and treats during my low moments.

To Siyang: Thank you for your presence as a steady light. No matter how many milestones you have reached in your own life, you never fail to cheer me up with the sweetest kindness (to the MAX!)

To my unfailing anchor: You are incredibly precious, a miracle of nearly a decade that continues to ground me when I lose my footing, even across four thousand miles.

Thank you, Mama, for your unconditional love, trust in my choices, and for always being my pillar of strength.

And finally, to Baba: This dissertation marks the end of a major chapter in my life, and wherever you are, I share this achievement with you. I hope you are still watching over me, as I know you always will be.

Abbreviations

ACC	Accusative
ASP	Aspect
AUX	Auxiliary
BA	Disposal marker
CL	Classifier
DAT	Dative
DE	Possessive/attributive particle
DUR	Durative
ERG	Ergative
FUT	Future
GEN	Genitive
LE	Sentence-final/Perfective particle
LOC	Locative
MEN	Plural marker for human nouns
NEG	Negation
NOM	Nominative
PART	Particle
PERF	Perfective
POSS	Possessive
PROG	Progressive
PST	Past
REL	Relativizer
RVC	Resultative Verb Compound
SUB	Subordinator

CHAPTER 1

Introduction

1.1 From functionalist observations to a formal semantic puzzle

A fundamental question in linguistics is how languages structure their referential systems to express definiteness and indefiniteness. While some languages systematically use articles (like the use of *the* and *a/an* in English), Mandarin allows for nouns without determiners (henceforth, **bare nouns**) and, consequently, seems to operate without a dedicated article system. This presents a puzzle: if Mandarin does not have a dedicated article system, how does it encode definiteness and indefiniteness?

Mandarin uses bare nouns like *shū* in (1) with both a definite and an indefinite interpretation. We render this in (1) by translating *shū* both as ‘the book’ and as ‘a book’.

- (1) wǒ yào qù mǎi **shū**.
I want go buy book
‘I want to go buy **the book/a book**.’

The interpretive flexibility of Mandarin bare nouns is well-established, but it makes it possible for Mandarin to resort to other forms in contexts in which languages like

2 Bare Nouns and Beyond

English use articles. The functionalist literature shows that Mandarin uses two forms in indefinite contexts and two forms in definite contexts. In indefinite contexts, the literature claims to find *yi* + classifier + noun (henceforth, **numeral-*yi***) alternating with bare nouns. In definite contexts, bare nouns are taken to alternate with demonstrative + classifier + noun (henceforth, **demonstratives**). In functionalist literature, numeral-*yi* and demonstratives function as article-like expressions and bleach their original numeral and demonstrative meanings (Chen, 2003, 2004; Diessel, 1999; Epstein, 1993; Givón, 1981, 1990; Greenberg, 1978; Huang, 1999; Li and Thompson, 1989; Lü, 1947). Relevant examples for numeral-*yi* and demonstratives are given in (2) and (3):

- (2) tā mǎi le **yì-běn shū**.
he buy PERF one-CL book
'He bought **a book**.' (Wright and Givón, 1987, 12)
- (3) yǒu yí-gè lièrén yǎng zhe yì-zhī gǒu. **zhè-zhī gǒu** hěn dǒngshì.
have one-CL hunter keep DUR one-CL dog. this-CL dog very intelligent
'There was a hunter who had a dog. **The dog** was very intelligent.'
(Chen, 2004, 1153)

In (2), numeral-*yi* appears in a canonical indefinite context, introducing a new discourse referent or a non-unique entity, much like an English indefinite article. In (3), the demonstrative appears in an anaphoric context and functions like an article.

While functionalist work focuses on the question of which forms are used in which contexts, the formal semantic literature on argument formation asks under which conditions nouns can function as arguments in definite and indefinite contexts. Based on examples like (1), this literature takes Mandarin nouns not to require additional formal support to function as arguments in either definite or indefinite contexts (Chierchia, 1998; Krifka, 2003; Dayal, 2004).

From the perspective of the functionalist literature on definiteness and indefiniteness, the formalist insight that Mandarin bare nouns do not require additional formal support to function as arguments poses no problem, as it does not lead to predictions that the functionalist literature itself does not make. However, the formal semantic literature on argument formation claims that the functionalist insight that bare nouns alternate with demonstratives/numeral-*yi* in (in)definite contexts does pose a challenge. Namely, if bare nouns do not require additional formal support in argument position, there is no clear reason to expect them to alternate with demonstratives or numeral-*yi*, contrary to the empirical claims of the functionalist literature. We conclude that

the functionalist literature’s empirical claims about alternations between bare nouns and demonstratives/numeral-*yi* pose a challenge for the formal semantic literature on argument formation. We refer to this challenge as the **alternation challenge**.

The goal of this dissertation is to address the alternation challenge by assessing whether the functionalist claims are empirically correct and, if so, by working out a formal semantic analysis that derives the facts. With this goal in mind, we present the theoretical background in the formal semantic literature on argument formation (Section 1.2), operationalize our goals in empirical and theoretical research questions (Section 1.3), and introduce our methodology and corresponding methodological research questions (Section 1.4). We conclude this introduction by delineating the argumentation of the dissertation (Section 1.5).

1.2 Theoretical background: the Kinds Approach and the Properties Approach

Even though formal semantic theories agree that Mandarin bare nouns do not require formal support to function as arguments, there is no consensus on how this capacity should be theoretically derived. Two main approaches can be distinguished: Krifka’s Properties Approach and Chierchia’s Kinds Approach. In this section, we work out the analyses proposed by the two approaches for argument formation in Mandarin and outline how these analyses fit the bigger cross-linguistic perspectives of the two approaches.

1.2.1 The formal analysis of Mandarin argument formation

The Properties Approach takes the view that Mandarin nouns start life as properties (type $\langle e, t \rangle$). They have to undergo a covert iota type-shift (ι) to appear as definite arguments (type $\langle e \rangle$) or a covert existential type-shift (\exists) to appear as indefinite arguments (type $\langle \langle e, t \rangle, t \rangle$). Meanwhile, the most recent version of the Kinds Approach (Jiang, 2020) posits that Mandarin nouns start life as kinds (type $\langle e \rangle$) and appear in definite contexts through an operation called Situation Restriction (SR), which takes a kind and returns the maximal member of that kind in a given situation. In indefinite contexts, Derived Kind Predication (DKP) comes into play. This operation takes a sortal predicate with a kind argument and returns the same predicate with existential quantification over instances of the kind.

4 Bare Nouns and Beyond

As we posited before, an alternation between bare nouns and demonstratives/numeral-*yi* would present a challenge for formal semantic theories of argument formation. We can now argue that this is indeed the case for both the Properties Approach and the Kinds Approach.

For the Properties Approach, the challenge is linked to the Blocking Principle, which states that the existence of an overt determiner τ blocks a bare noun χ from covertly acquiring the meaning of $\tau\chi$. The upshot of this principle is that the Properties Approach cannot allow nouns preceded by determiners and bare nouns to convey the same meaning in argument position. That is, if a bare noun χ can express the meaning of $\iota\chi$ or $\exists\chi$, there should not be a determiner with the meaning ι or \exists . Based on Krifka's assumption that Mandarin uses bare nouns in argument position in regular (in)definite contexts, there is no place for nouns preceded by demonstratives/numeral-*yi* to appear in the same positions. This means that Krifka relegates demonstratives to contexts that revolve around typical ingredients of demonstratives (e.g., deixis) and that he restricts the use of numeral-*yi* to contexts in which cardinality matters, i.e., contexts in which we typically find numerals. Figure 1.1 visualizes these predictions.

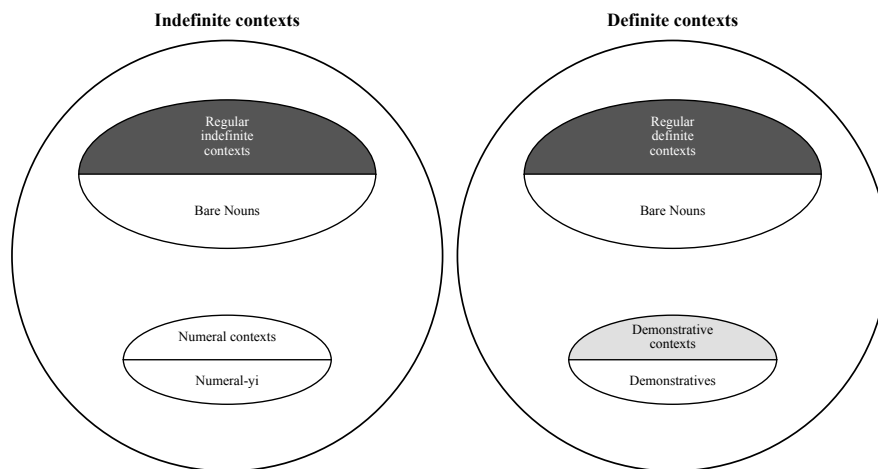


Figure 1.1: Predictions by the Properties Approach on the Distribution of Bare Nouns, Numeral-*yi*, and Demonstratives.

Figure 1.1 represents the predictions of the Properties Approach under Krifka's assumption that bare nouns are used in argument position in regular definite and indefinite contexts. According to this assumption, there is no room for alternations between bare nouns and demonstratives/numeral-*yi* in regular (in)definite contexts, and

demonstratives/numeral-*yi* are restricted to demonstrative and numeral contexts.

The alternation challenge comes about differently in the Kinds Approach. DKP and SR are, strictly speaking, not type-shifts, and therefore, they are not sensitive to the Blocking Principle. As such, no mechanism in the Kinds Approach blocks bare nouns from alternating with nouns preceded by demonstratives/numeral-*yi* in (in)definite contexts. However, it remains unclear why such an alternation would exist in the first place. Indeed, within the Kinds Approach, Mandarin is the prototype of a language that lacks all functional pressure for demonstratives/numeral-*yi* to develop into markers of argumenthood in (in)definite contexts. Mandarin nouns are argumental from the start, so there is no reason for Mandarin to develop determiners to mark argumenthood. Consequently, the net predictions of the Kinds Approach are the same as those of the Properties Approach, as discussed above and represented in Figure 1.1. There can be no alternations between bare nouns and demonstratives/numeral-*yi* in regular (in)definite contexts, and the latter are restricted to demonstrative and numeral contexts.

Based on the preceding discussion, we conclude that alternations between bare nouns and demonstratives/numeral-*yi* constitute a challenge for the Properties and Kinds Approaches. In their latest analyses of Mandarin, both approaches only allow bare nouns to appear in regular (in)definite contexts and relegate demonstratives/numeral-*yi* to demonstrative and numeral contexts.

1.2.2 Cross-linguistic perspectives

A core assumption of the Kinds and the Properties Approaches is that there is a universal system of argument formation that Mandarin conforms to and informs us about. As such, the analysis of argument formation in Mandarin can (and should) be seen as part of the broader cross-linguistic views on argument formation that the two approaches argue for.

The two approaches share a basic architecture that consists of the type-shifting framework presented by Partee (1987) and further elaborated by Chierchia (1998). In this framework, nominal expressions can correspond to different semantic types (entities (type $\langle e \rangle$), predicates (type $\langle e, t \rangle$), and quantifiers (type $\langle \langle e, t \rangle, t \rangle$)), and a number of basic type-shifting operations allow them to shift between these types. We already saw two of these type-shifts in our discussion of the Properties Approach's analysis of argument formation in Mandarin, *viz.*, the iota type-shift (ι) and the existential type-shift (\exists). The other basic type-shift relevant for our discussion is the down type-shift

6 Bare Nouns and Beyond

(¹). This shift takes a non-singular property and returns its corresponding kind, if defined.

The two approaches have a shared understanding of what “regular” definite and indefinite contexts are, *viz.*, those having a semantics that involve or are equivalent to the result of an iota type-shift (for definites) or an existential type-shift (for indefinites). Both approaches also share the assumption that articles like *the* and *a/an* in English and other article languages are the overt realizations of these type-shifts. Finally, both approaches also follow the Blocking Principle, which states that the existence of an overt determiner with the semantics of one of the basic type-shifts blocks this type-shift from applying covertly.

The Properties Approach adds one assumption to the basic architecture we have just sketched, *viz.*, that nouns always start life as properties, independent of the language. In the Properties Approach, Mandarin does not differ from other languages: Mandarin nouns start life as properties, and even though Krifka assumes that Mandarin does not have articles, there is nothing intrinsic about the Properties Approach’s take on Mandarin that would prevent it from developing articles in the same way English has.

The Kinds Approach adds two assumptions to the basic architecture outlined above. The first is that nouns can start life as properties (type $\langle e, t \rangle$) or as kinds (type $\langle e \rangle$), depending on the language. In Chierchia (1998), this assumption led to a ternary division of languages known as the Nominal Mapping Parameter: languages with nouns that start life as properties, languages with nouns that start life as kinds, and languages with both types of nouns. In more recent work, this ternary typology has been reduced to a binary one (Chierchia, 2010; Jiang, 2020); this is the one we adopt. The binary typology distinguishes between languages in which nouns start life as properties and those in which nouns start life as kinds. English is assumed to be of the former type, and Mandarin is assumed to be of the latter type. Thus, in the Kinds Approach, Mandarin is considered to be different from languages in which nouns start life as properties. From the perspective of argument formation, the main difference is that languages with property-nouns may have (but need not develop) articles, whereas languages like Mandarin are predicted not to develop articles in view of the argumental status of their nouns.

The second assumption that the Kinds Approach adds to the basic architecture noted above is that type-shifts are ranked. Chierchia (1998) argued that type-shifts leading to simpler types should be preferred over those leading to more complex types. This led him to posit that the down type-shift should be preferred over the existential

type-shift ($\exists < \cap$). Later, Dayal (2004) generalized Chierchia's logic and proposed ranking the iota type-shift on a par with the down type-shift, leading to the ranking system that is now considered standard in the Kinds Approach: $\exists < \{\iota, \cap\}$. This is the version of the ranking we adopt.

1.2.3 The alternation challenge: summary

This section introduced Krifka's Properties Approach and Chierchia's Kinds Approach, the two main formal semantic approaches to argument formation in Mandarin. We also sketched the role Mandarin plays in the broader cross-linguistic claims the two approaches make.

We argued that both approaches predict that Mandarin does not allow for alternations between bare nouns and demonstratives/numeral-*yi* in definite and indefinite contexts. The lack of alternations in the Properties Approach is linked to the Blocking Principle, which prevents bare nouns and demonstratives/numeral-*yi* from having the same semantics. Meanwhile, the predicted lack of alternations in the Kinds Approach follows from the assumption that Mandarin nouns are argumental from the start, precluding all functional pressure for articles to develop.

From the broader cross-linguistic perspectives of the two approaches, we established that the Properties Approach considers Mandarin to be like any other language in that its nouns start life as properties and that it may (but need not) develop articles. This is different for the Kinds Approach, for which Mandarin occupies a special place as a language in which nouns start life as kinds and articles are not expected to develop. The two approaches agree on the following assumptions: (i) definite and indefinite contexts involve or are equivalent to the result of an iota or an existential type-shift, (ii) articles have these type-shifts as their semantics and block them from applying covertly, and (iii) English *the* and *a/an* are prototypical articles. Finally, we drew attention to a second difference between the two approaches: whereas the Properties Approach imposes no ranking on type-shifts, the Kinds Approach ranks the existential type-shift below the iota and the down type-shift ($\exists < \{\iota, \cap\}$).

We have now established that alternations between bare nouns and demonstratives/numeral-*yi* would constitute a challenge both for the Properties Approach and the Kinds Approach. If the alternation claims of the functionalist literature are correct, we cannot but reconsider the analyses proposed by the formal semantics literature on argument formation for Mandarin. We have also laid the foundation for our argumentation that (as we will see) heavily relies on cross-linguistic comparison. As such, the

similarities and differences between the broader cross-linguistic perspectives of the two approaches must be considered.

1.3 Guiding empirical and theoretical questions

With our theoretical background in place, we can operationalize the dissertation goal set in Section 1.1 in terms of two overarching research questions:

1. Empirically, how does Mandarin encode definiteness and indefiniteness? Do we only find bare nouns, or do we also find numeral-*yi* and/or demonstratives in indefinite and definite contexts? And if we find both, what is the division of labor?
2. Theoretically, what do our empirical findings mean for formal semantic approaches to argument formation? Do we need to develop a fundamentally different alternative to existing approaches like the Kinds and Properties Approaches, or can we account for the alternations and derive the division of labor within at least one of these approaches?

With these questions in place, we are almost ready to provide the argumentation of this dissertation. However, we are missing one crucial element, namely, the methodology used to ground our empirical findings. This is the topic of Section 1.4. Then, once the methodology is in place, Section 1.5 will explain how this dissertation connects methodology, data, and theory.

1.4 Methodology: a translation corpus approach

The methodological challenge faced in this dissertation is that we want to study how definiteness and indefiniteness are marked in a language for which we cannot presume that it has a systematic formal paradigm in place that distinguishes between the two. Examples like (1) raise the question of how we can tell whether we are dealing with a definite or an indefinite context. Examples like those in (2) and (3) are even trickier: How can we know whether we are dealing with regular indefinite and definite contexts rather than numeral and demonstrative contexts? The key, we argue, lies in our theoretical background and the link it allows us to make between the Mandarin examples in (1), (2), and (3) on the one hand and their translations on the other hand. We remind the reader that the Properties and Kinds Approaches both assume that English *the* and

a/an are in a one-to-one relation with regular definite and indefinite contexts. Based on this assumption, we can use the translation equivalence with English to establish what types of contexts we are dealing with for any language in general and for Mandarin in particular. We conclude that a translation approach constitutes a viable route that is worth exploring; this is why we adopt it in this dissertation.

Adopting a translation approach leaves open the option to work with introspection, translation experiments, or translation corpora. In this dissertation, we opt for corpora. This means we take a data-driven approach that allows for a random selection of data that is not guided by our intuitions, as in an introspection approach, nor by a preset selection of contexts, as is typical in an experimental approach.

For our corpus, we initially drew on J.K. Rowling's *Harry Potter and the Philosopher's Stone* and its Mandarin translation. We chose this corpus for two reasons. The first is that it is a relatively recent novel with, accordingly, a relatively recent translation. The second is the extensibility of the corpus. As will become clear, our work on Mandarin brings us beyond a comparison between Mandarin and English, and the fact that *Harry Potter and the Philosopher's Stone* has been translated into 80+ languages makes our initial corpus easily extendible.

To the best of our knowledge, we are the first to apply a translation corpus approach to the study of argument formation in Mandarin. As a consequence, we could not help but stumble on methodological questions. Thus, part of our work will also aim to answer the following guiding methodological research question that complements the empirical and theoretical ones we introduced in Section 1.3:

3. Methodologically, how can we ensure that a translation corpus approach provides reliable data that allow us to answer our empirical questions?

This guiding methodological question can be subdivided into four subquestions. The first has to do with the risk of taking *the* and *a/an* as a starting point: a focus on translations of NPs preceded by *the* and *a/an* will allow us to identify the expressions that are used as translations but will not give us any insight into how these expressions are used in other contexts. The risk we run, then, is that of looking at these expressions only through the lens of English *the* and *a/an*. The question that raises is how we can make sure we do justice to Mandarin expressions *per se*, as opposed to treating them only as translations of NPs preceded by *the* and *a/an*. We refer to this question as Subquestion 3.1.

The second subquestion has to do with the inherent limitations of the use of a corpus. Even though corpora allow us to avoid the biases associated with introspection

and experiments, a random corpus selection might still have its own biases in that certain context types may be over- or underrepresented. Indeed, if we find that English *a/an* is often translated by numeral-*yi*, how can we make sure that the contexts we are looking into do not straightforwardly invite the use of a numeral in the first place? The more general version of this question is Subquestion 3.2: How can we make sure that the potential biases of our random corpus selection do not limit the scope of the empirical answers our translation corpus approach can provide?

The third and fourth subquestions are related to two oft-cited limitations of translation corpus data, namely, that translations might not be representative of their target languages (see Le Bruyn et al., 2022, and references therein) and that meaning differences can exist between originals and translations (see Le Bruyn et al., 2024, and references therein). Accordingly, Subquestion 3.3 is, “How can we make sure that translations are sufficiently representative of their target languages?” Moreover, Subquestion 3.4 asks, “How can we ensure that meaning differences between originals and translations do not preclude drawing valid conclusions based on translated data?”

Regarding Subquestions 3.2 to 3.4, we submit that, ultimately, the limitations of one method can be properly addressed only by triangulation through another method (Le Bruyn et al., 2022, 2024; Le Bruyn and de Swart, 2024). However, given that we are the first to apply a translation corpus approach to Mandarin argument formation, combining it with other methodologies would lead us too far. Consequently, we will present ways to address the limitations of Subquestion 3.2 within translation corpus research itself. Building on the insight that the potential ambiguity of examples like (1) is resolved in context, de Swart (2020) argued that multilingual translation corpora can overcome the limitations of a comparison between two languages. For this strategy to work, we need to select languages that complement each other in their grammatical set-up. As will become clear in our argumentation overview in Section 1.5, the inclusion of data from languages other than English and Mandarin plays a pivotal role. The investigation of Hindi, Russian, German, Spanish, and Hebrew adds languages to the multilingual dataset that have been argued to lack articles (like Mandarin), languages with a full-fledged article system (like English), and a language with a definite article but no indefinite article. The unique position of Mandarin within this multilingual comparison provides the starting point of the theoretical reflection.

1.5 The argumentation of this dissertation

In this section, we provide the argumentation of this dissertation as a chapter-by-chapter overview. Three chapters in this dissertation (Chapters 3, 4 and 6) have been published as independent co-authored papers, each of which was optimized for a specific venue and presents the shared views of its authors. This dissertation includes each paper as it was published, and we use this section to show how the research reported in the papers should be integrated into this dissertation’s overarching argumentation.

Chapter 2: A Corpus Investigation of the Mandarin Referential System

In Chapter 2, we employed our translation corpus approach to provide a preliminary answer to our guiding empirical question concerning how Mandarin encodes definiteness and indefiniteness. As indicated in Section 1.4, we did this by examining *Harry Potter and the Philosopher’s Stone* (henceforth HP) and its Mandarin translation to check how English NPs introduced by *the* and *alan* are translated in Mandarin. In the same chapter, we organized the design and results of the study around a series of hypotheses based on the functionalist literature on definiteness and indefiniteness and on the formal semantics literature on argument formation. Here, we focus on the overall design and results on the one hand and on the main methodological and empirical conclusions that we draw on the other hand.

When designing the study, we addressed two of our methodological subquestions. The first is Subquestion 3.1, *viz.*, “How can we make sure that we do justice to Mandarin expressions in their own right and not only as translations of NPs preceded by *the* and *alan*?” Our answer is inspired by Le Bruyn et al. (2024): to make sure that we did justice to all uses of the Mandarin expressions that appear as translations of NPs preceded by *the* and *alan*, we looked at translations not only of the latter but also of all referential expressions. By doing so, we ensured that we accessed the full referential system of Mandarin and that we did not cut off the referential expressions that appeared as translations of NPs preceded by *the* and *alan* from their uses in other referential contexts. Extending our attention to all referential expressions in English and Mandarin also allowed us to address methodological Subquestion 3.3, *viz.*, “How can we argue that translation data are representative of the target language?” Our answer is again inspired by Le Bruyn et al. (2024): by looking at all referential expressions, we were able to build independent checks of target language representativeness into

our design. These checks concern data patterns that are independent of the core definiteness and indefiniteness data that we study: if we found that these data patterns were in line with predictions in the literature, we would have independent evidence for the target language representativeness of our translated Mandarin data. We refer to the Mandarin expressions included in these data patterns as benchmark categories. As we decided to focus on all referential expressions, we could not take into account our full corpus. Accordingly, the study described in Chapter 2 focuses on one chapter of HP and its translation, totaling over 1200 English referential expressions and their translations into Mandarin.

The findings from Chapter 2 are as follows. First, our benchmark categories behaved as expected. Second, NPs preceded by *the* are translated by bare nouns (81%) and demonstratives (14%), and other Mandarin expressions do not have a role of significance to play in the translations. Third, NPs preceded by *alan* are translated by bare nouns (26%) and numeral-*yi* (66%), and, again, other Mandarin expressions do not have a role of significance to play. Fourth, next to their appearance in the translations of NPs preceded by *the* and *alan*, demonstratives and numeral-*yi* mainly appear as translations of demonstratives and the numeral-one.

Methodologically, the above findings lead to two conclusions. First, a translation corpus approach that focuses on all referential expressions from the source language allowed us to do justice to the referential expressions in the target language in their own right and not only as translations. In our particular case, this translation corpus strategy allowed us to establish that Mandarin demonstratives and numerals function in the same way as demonstratives and numerals in English but have an extended use in definite and indefinite contexts. Second, the fact that our benchmark categories behaved as expected means that our translated referential data are representative of the target language. These conclusions provide our answers to methodological Subquestions 3.1 and 3.3. First, by looking at a broader range of data from the source language, we do justice to target language expressions in their own right; second, by including benchmark categories, we can establish the target language representativeness of our translation corpus data.

Empirically, our findings show that, as far as our corpus is concerned, demonstratives and numeral-*yi* are used in demonstrative and numeral contexts and alongside bare nouns in definite and indefinite contexts. Mandarin demonstratives and numerals thus function as demonstratives and numerals in English, except that they also occur in definite and indefinite contexts. This means that, all other things being equal, the claims of the functionalist literature about alternations of bare nouns and

demonstratives/numeral-*yi* in definite and indefinite contexts are on the right track. There is an important caveat, though, that we need to confirm that the contexts from our corpus in which English uses NPs preceded by *the* or *a/an* are not biased such that we cannot extend the scope of our conclusion beyond our corpus.

Chapter 3: “Articleless” Languages are Not Created Equal

Chapter 3 was originally published with the title “Articleless” languages are not created equal in the proceedings of Sinn und Bedeutung 27 and is joint work with Shravan Patil, Daria Seres, Olga Borik and Bert Le Bruyn. The original paper is as an evaluation of Dayal (2004). In line with the argumentation of this dissertation, we focus on the answers the data and methodological choices in the paper provide to our empirical and methodological research questions.

The challenge we identified in Chapter 2 is linked to methodological Subquestion 3.2: “How can we argue that there are no problematic biases in a randomly selected corpus?” The answer we propose for the corpus we used in Chapter 2 is based on the following insight: if there are biases that run counter to the conclusion of alternation that suggests itself for our corpus, they would be such that demonstratives and numeral-*yi* appear as translations of NPs preceded by *the* and *a/an* because the contexts in which the latter appear invite the use of bare nouns, demonstratives, and numerals on their standard demonstrative and numeral uses. To examine whether the relevant biases are present in our selection of contexts, we can check how the same contexts are translated in other languages that have been argued not to have articles. If the use of demonstratives or numeral-one in the Mandarin translation is driven by contexts that invite the use of demonstratives and numeral-one on their standard demonstrative and numeral uses, we would expect similar patterns to appear in other articleless languages, all other things being equal.

The corpus study we report on in Chapter 3 considers the same contexts as in Chapter 2 but focuses on those that include NPs preceded by *the* and *a/an* in the English original. At the same time, the scope of the study is broadened from Mandarin translations to Russian and Hindi translations. We chose these languages because they are the three “articleless” languages that Veneeta Dayal treats in her seminal paper from 2004. For Hindi, Dayal claims that bare plurals can freely appear in definite and indefinite contexts but that bare singulars can only appear in definite contexts. For a singular noun to appear in indefinite contexts, it would either need to take numeral-one or appear in a pseudo-incorporated construction. Dayal makes the same claim

for Russian, but its validity has been questioned by several authors who have argued that Russian bare nouns (both singular and plural nouns) can freely appear in definite and indefinite contexts (Bronnikov, 2006; Šimík and Demian, 2020; Seres and Borik, 2018). Finally, for Mandarin, Dayal claims that bare nouns freely appear in definite and indefinite contexts. The empirical claims of the functionalist literature argue against Dayal, suggesting that demonstratives and numeral-*yi* play a role in definite and indefinite contexts.

If Dayal is right about Hindi, Russian, and Mandarin, we expect to find that the frequency of Mandarin demonstratives in our corpus is not significantly higher than that of demonstratives in Russian and Hindi and that the frequency of numeral-*yi* is significantly lower than that of its counterparts in Russian and Hindi. If Dayal is right about Hindi and Mandarin—and Bronnikov (2006) and others are right about Russian—our expectations for demonstratives remain unchanged, but our expectations for numeral-*yi* change in that we do not expect its frequency to be significantly higher than in Russian. Finally, if Dayal is correct about Hindi, Bronnikov (2006) and others are right about Russian, and the functionalist literature is correct about Mandarin, we expect to find that the frequency of demonstratives in Mandarin is significantly higher than that in Hindi and Russian and that the frequency of numeral-*yi* is significantly higher than that in Russian.

The most striking empirical findings of the research reported in Chapter 3 are as follows. For Russian, bare singulars freely appear in definite and indefinite contexts, with both types of contexts having 80% or more of bare nouns. For Hindi, bare singulars also make up more than 80% of referential expressions in definite contexts but only around 40% of referential expressions in indefinite contexts, where they have the same frequency as NPs preceded by numeral-one. We also found that demonstratives qualify as the most important alternatives to bare nouns both in Russian and Hindi, accounting for around 2% and 7% of referential expressions in definite contexts, respectively. To further interpret these quantitative findings, we ran Fisher's exact tests to test the significance of the differences in proportions of bare nouns, demonstratives, and numeral-one in definite and indefinite contexts across the three languages. The tests reveal that Hindi and Russian do not significantly differ in their proportion of demonstratives and bare nouns in definite contexts, and the same holds for Hindi and Mandarin. However, we did find a significant difference between Russian and Mandarin. For indefinite contexts, Fisher's exact tests reveal significant differences between the three language pairs.

The above findings led us to draw two conclusions for the argumentation presented

in this dissertation. The first concerns the alternation we found between bare nouns and numeral-*yi* in Chapter 2. The comparison between Mandarin and Russian shows that there is no reason to assume that the indefinite contexts in our corpus from Chapter 2 are biased in that they would invite the use of both bare nouns and numerals. The difference in the use of bare nouns and numeral-one between Russian and Hindi does not change this: the discussion in the literature between Dayal, Bronnikov, and others already suggests a difference between the use of numeral-one in Russian and Hindi. Thus, we conclude that the scope of our empirical findings about bare nouns and numeral-*yi* in indefinite contexts in Chapter 2 can be generalized beyond our corpus and that the alternation that the functionalist literature claims to exist between bare nouns and numeral-*yi* in indefinite contexts is real.

The second conclusion regards the alternation we found between bare nouns and demonstratives in the corpus we used in Chapter 2. Specifically, our cross-linguistic findings provide no grounds to definitively dismiss the possibility that the contexts in our corpus are biased towards inviting both bare nouns and demonstratives. Definitive evidence would have shown significant differences between Russian and Mandarin and between Hindi and Mandarin, but we did not find such differences. At the same time, we note that demonstratives are more frequent in Mandarin than in Russian and Hindi and that our cross-linguistic study might have been too small to reveal significant differences. Therefore, our cross-linguistic study does not allow us to generalize Chapter 2's empirical findings about definite contexts beyond our corpus. At the same time, we must acknowledge that we do not have definitive evidence to question the alternation that the functionalist literature claims exists between bare nouns and demonstratives.

On the methodological side, the conclusion we draw from the research reported in Chapter 3 is strongly intertwined with our empirical conclusions. Indeed, the reason we can draw the above empirical conclusions is that the cross-linguistic design of our study allowed us to generalize some of the conclusions we reached in Chapter 2 beyond our corpus. Thus, our answer to methodological Subquestion 3.2 about the generalizability of corpus-based findings in translation corpus research is as follows: within a translation corpus approach, a cross-linguistic design provides a straightforward way to control for problematic biases in translated data.

The research presented in Chapters 2 and 3 allows us to make significant headway both empirically and methodologically. Empirically, it establishes that the alternations that the functionalist literature claims exist between bare nouns and numeral-*yi* are real. Methodologically, the research shows that a design in which we consider a broad

range of expressions in the source language—and in which we adopt a cross-linguistic rather than a contrastive perspective—allows us to minimize or even overcome the limitations addressed in methodological Subquestions 3.1–3.3.

The confirmation of at least some of the alternation claims of the functionalist literature invites us to reflect on how alternation data can be analyzed from the perspective of the formal literature on argument formation. However, with the cross-linguistic turn our research has taken, limiting our reflection to Mandarin alone would neglect some of our data. Consequently, the next step is to consider the data from Mandarin along with those from Hindi and Russian and discuss how to best proceed from the perspective of the two formal semantic frameworks we introduced in Section 1.2, *viz.*, the Kinds and Properties Approaches. We do this based on the research described in Chapter 4. We also use this research to address our last methodological subquestion, *viz.*, Subquestion 3.4 regarding the justification of the assumption of meaning stability that underlies translation corpus research.

Chapter 4: The Theory of Argument Formation: between Kinds and Properties

Chapter 4 was originally published with the title The theory of argument formation: between kinds and properties in the proceedings of SALT 33 and is joint work with Shravani Patil, Hagay Schurr, Daria Seres, Olga Borik and Bert Le Bruyn. The original paper is an exploration of the options the Kinds and Properties Approaches have to account for data from a typologically broad range of languages. So that it fits the argumentation of this dissertation, we slightly adapt the exploratory theoretical perspective and restrict it to Mandarin, Hindi, and Russian. We use the data from other languages (Spanish, German, and Hebrew) to address methodological Subquestion 3.4.

The research reported in Chapter 4 utilizes the same data as in Chapters 2 and 3 but extends them one last time with translations to Spanish, German, and Hebrew. In line with the argumentation of this dissertation, we build on the theoretical exploration in the original paper to provide a preliminary answer to our guiding theoretical question, *viz.*, “What do our empirical findings mean for formal semantic theories of argument formation?” The data from Russian, Hindi, and Mandarin are of central concern for this theoretical question. We keep the data from Spanish, German, and Hebrew separate to address methodological Subquestion 3.4. In what follows, we first present the methodological step we took based on the research described in the chapter before

turning to the theoretical step. Methodological Subquestion 3.4 targets the validity of the basic assumption of translation corpus research in linguistics, *viz.*, that meaning is kept constant across translations. This might seem like an obvious assumption for readers who are not used to working with translations; however, within translation corpus research, it is clear that this assumption cannot be taken for granted and should, where possible, be argued for. We submit that certain meaning components are more prone than others to meaning change between translations and that the assumption of meaning stability should be justified not in general but for the phenomena one studies. In our case, these are definiteness and indefiniteness.

Ultimately, meaning stability should be argued for at the level of individual translations. This would, however, lead us far beyond the scope of this dissertation and would probably require a triangulation strategy. As a compromise, we argue for the stability of definiteness and indefiniteness based on a rationale that exploits cross-linguistic comparison. Namely, the relative stability of definiteness and indefiniteness across translations can be estimated based on languages for which the literature conveys definiteness and indefiniteness consistently. If we find that the translation patterns in these languages are in line with what we expect based on the literature, we have reasonable grounds for our assumption about meaning stability of definiteness and indefiniteness. The core empirical findings for Spanish, German, and Hebrew are as follows. Spanish and German typically use indefinite articles to translate NPs preceded by *alan* (around 80% of the translations) and definite articles to translate NPs preceded by *the* (around 80% of the translations). Hebrew typically uses bare nouns for the former (around 90% of the translations) and the definite article or a construct state for the latter (around 80% of the translations). These data patterns are in line with the literature on these languages (see, e.g., Espinal and McNally, 2011, on Spanish; Löbner, 2011, on German; and Doron, 2003, on Hebrew). Consequently, we conclude that definiteness and indefiniteness are extremely stable across translations and that there is no reason to expect the translations we rely on to misrepresent the meaning of the original. Turning now to our theoretical exploration of the Mandarin, Hindi, and Russian data, we argue that neither the Kinds Approach nor the Properties Approach can, in their current form, account for the data of each of these languages. For the Kinds Approach, Mandarin and Russian constitute a problem, and for the Properties Approach, Mandarin and Hindi constitute a problem. Specifically, the fact that Russian bare singulars freely appear in indefinite contexts is problematic for the Kinds Approach. Indeed, given that the Kinds Approach assumes that Russian nouns start life as properties, that the down type-shift is unavailable for singular nouns, and that

the iota type-shift is ranked above the existential type-shift, Russian bare singulars should only appear in definite contexts. Meanwhile, the Properties Approach does not assume that type-shifts are ranked and, thus, correctly predicts that bare singulars in Russian can appear in indefinite contexts. For Hindi, the Kinds Approach makes accurate predictions in that the ranking of type-shifts allows bare singulars to freely occur in definite contexts but not in indefinite contexts. Within this approach, the bare nouns we find in indefinite contexts are taken to be pseudo-incorporated. Meanwhile, the Properties Approach assumes that Hindi does not have articles and, consequently, fails to predict that bare singulars are not free to occur in indefinite contexts. Finally, in Chapter 3, we established that the alternation challenge is real at least for indefinite contexts and, thus, that both approaches face this challenge for Mandarin. The above overview shows that each of the approaches, in its current form, fails to account for the Mandarin data and for the data of at least one other language. In an optimal scenario, we would develop an account that covers not only the Mandarin data but also the Russian and Hindi data. Our guiding theoretical question asks whether we need to develop a fundamentally different alternative to existing argument formation approaches to derive the data patterns that we find. We respond to this question by exploring the options available within existing approaches. However, doing so for two theoretical approaches would lead us beyond the scope of this dissertation. Therefore, we focus on the Properties Approach and propose several hypotheses that (for Russian and Hindi) rely on existing analyses while holding the promise of deriving the data patterns we find in the three languages. For Russian, the Properties Approach does not need to be altered, as it straightforwardly derives the free occurrence of bare nouns in definite and indefinite contexts found in our data. For Hindi, the Properties Approach needs to be adapted, but the adaptation does not require us to fundamentally rethink the approach. We only need to change the assumption that Hindi does not have an indefinite article. The comparison between our Russian and Hindi data shows that there is a significant difference in the use of numeral-one between the two languages, which provides sufficient grounds for hypothesizing that numeral-one in Hindi functions as an indefinite article. With this hypothesis in place, only one challenge remains for the Properties Approach in Hindi. Namely, we also found bare nouns as translations of NPs preceded by *a/an* even though the Blocking Principle should disallow this state of affairs. This challenge can be overcome, though, by following Dayal in hypothesizing that bare nouns in indefinite contexts in Hindi do not function as regular arguments but are pseudo-incorporated. There is nothing in the Properties Approach that refutes Dayal's hypothesis. With Russian and Hindi covered, we can turn to the

question of how to deal with the alternation challenge in Mandarin. The key hypothesis we propose to account for the alternation we find in indefinite contexts is that bare nouns can only occur in pseudo-incorporated positions. Then, numeral-*yi* can be analyzed as an indefinite article and taken to block bare nouns from appearing in regular argument positions. For bare nouns and demonstratives in definite contexts, we have no definitive grounds to accept or reject the notion that they alternate. Consequently, we propose two hypotheses. The first is that there is a real alternation and that it revolves around the weak/strong definiteness distinction, as argued by Jenks (2018). Within the Properties Approach, this distinction can be modeled as involving two versions of the iota type-shift and would be compatible with the Blocking Principle. The second hypothesis is that there is no real alternation and that the demonstratives we find in definite contexts are regular demonstratives. According to these hypotheses, the Properties Approach accounts for our definiteness and indefiniteness data from Russian, Hindi, and Mandarin without having to reject its core assumptions or the Blocking Principle. We repeat that we do not deny that alternative hypotheses can be formulated to account for our data in the Kinds Approach but that this would lead us beyond the scope of this dissertation. At the same time, we see challenges for the Kinds Approach, as detailed in Chapter 4. Methodologically, the research reviewed in Chapter 4 allowed us to address methodological Subquestion 3.4 and conclude that a cross-linguistic methodology allows us to confirm that the meaning components we research on—definiteness and indefiniteness—are extremely stable across translations and, consequently, that we can safely rely on the assumption of meaning stability in our research. When we add this conclusion to those provided in Chapters 2 and 3, a broader conclusion emerges, namely, that the definiteness and indefiniteness data that we uncovered by looking at translations of NPs introduced by *the* and *alan* give us a firm understanding of definiteness and indefiniteness in Mandarin. We thus take our translation corpus approach to be fully argued for, and we deploy it in the rest of our argumentation as a received approach. Theoretically, the research discussed in Chapter 4 allowed us to formulate a number of hypotheses that hold the promise of deriving our definiteness and indefiniteness data from Russian, Hindi, and Mandarin in the Properties Approach. The fact that these hypotheses do not require any major rethinking of the Properties Approach allows for a (very tentative) first answer to our guiding theoretical research question, *viz.*, that we do not have to propose a fundamentally different alternative to existing formal semantic theories of argument formation to derive the data patterns we find. This answer is tentative because some of the hypotheses still need to be argued for, and the analyses that go with them still need to

be worked out. We submit that the hypotheses we proposed for Hindi are sufficiently worked out in the literature. However, the same does not hold for the hypotheses we proposed for Mandarin, and this is the issue we tackle in Chapters 5, 6 and 7. Chapter 5 focuses on indefiniteness, while Chapters 6 and 7 focus on definiteness.

Chapter 5: Fine-Tuning the Property-Based Analysis: the Indefinite Domain

Based on the research reported in Chapter 4, we formulated a hypothesis that could allow us to derive the alternation we find between numeral-*yi* and bare nouns in indefinite contexts within the Properties Approach, numeral-*yi*, that bare nouns in indefinite contexts are pseudo-incorporated. In Chapter 5, we test this hypothesis and conclude that there is reason to distinguish between regular arguments and pseudo-incorporation in indefinite contexts and to restrict bare nouns to pseudo-incorporation positions. Consequently, we develop a new analysis of pseudo-incorporation in Mandarin that, when combined with an article analysis of numeral-*yi*, derives the alternation between bare nouns and numeral-*yi* in indefinite contexts within the Properties Approach.

The empirical challenge we tackle in this chapter is that the pseudo-incorporated status of nouns is typically argued for based on a series of morpho-syntactic properties (e.g., number, case, word order) that have no relevant counterparts in Mandarin (Borik and Gehrke, 2015). Building on existing work on pseudo-incorporation in other languages and on the work of Huang (2015) and Luo (2022) for Mandarin, we propose the typicality of the relation between verbs and nouns as a semantic criterion to argue for the existence of pseudo-incorporation in Mandarin. Huang (2015) and Luo (2022) argued that pseudo-incorporation exists in Mandarin and that it requires a typicality relation between verbs and nouns. However, they restricted pseudo-incorporation to a number of more or less fixed expressions, whereas we argue that the same typicality relation can insightfully be used to distinguish indefinite contexts in which bare nouns are allowed from those in which the use of numeral-*yi* imposes itself.

We propose that the pseudo-incorporation of nouns is only possible if the nouns stand in a typical relation to their verbs (e.g., wear suit, read book). If bare nouns can only appear in indefinite contexts if they are pseudo-incorporated, then we predict that they will only occur in combination with verbs that they stand in a typical relation to.

To check whether bare nouns only appear with verbs that they stand in a typical relation to, we used a translation corpus approach based on our full HP corpus. We

observed the Mandarin translations of the occurrences of NPs preceded by *a* that occur in object position and restrict them further to translations by numeral-*yi* and bare nouns that occur in object position themselves ($n = 154$). The results confirm that bare nouns only occur as objects of verbs that they stand in a typical relation to. These results led us to conclude that bare nouns in Mandarin are pseudo-incorporated.

In this chapter, we tackle the theoretical challenge of designing an analysis of pseudo-incorporation that is compatible with the Properties Approach, that captures the typicality requirement, and that allows us to distinguish pseudo-incorporation in Mandarin from pseudo-incorporation in languages with stricter or looser restrictions on the verb-noun combinations allowing for pseudo-incorporation.

Our analysis method combines insights from Espinal and McNally (2011), Dayal (2011) and Le Bruyn et al. (2016). Its compatibility with the Properties Approach is guaranteed through the assumption that pseudo-incorporated nouns are of type $\langle e, t \rangle$ and existential import is not obtained through type-shifting of the noun but through a lexical rule that applies to verbs. The typicality requirement is captured by making the lexical rule sensitive to the semantics of the verb and to the QUALIA structure of the noun. Our analysis differs from those that rely on a special mode of composition that we take to be reserved for languages with no restrictions on the verbs and nouns that allow for pseudo-incorporation. Because our analysis relies on a lexical rule, it is flexible enough to deal with languages that have stricter or looser restrictions on the verbs and/or nouns that allow for pseudo-incorporation.

The empirical and methodological work in this dissertation, as described at the end of Chapter 5, allows us to conclude that the pseudo-incorporation hypothesis we proposed based on the research reviewed in Chapter 4 is supported. In turn, we can draw the following empirical conclusions:

- NPs preceded by numeral-*yi* and bare nouns alternate in indefinite contexts;
- The bare nouns that we found in indefinite contexts are pseudo-incorporated and stand in a typicality relation to the verbs they combine with.

Moreover, our theoretical work allows us to conclude that the Properties Approach can derive these facts by making the following assumptions:

- Numeral-*yi* functions as an indefinite article and blocks bare nouns from appearing in regular argument position in indefinite contexts;
- Mandarin pseudo-incorporated nouns are of type $\langle e, t \rangle$, and existential import is obtained through a lexical rule applied to verbs;

- This lexical rule is sensitive to the presence or absence of a typicality relation between verbs and nouns.

Chapter 6: Translation Mining: Definiteness across Languages (A Reply to Jenks 2018)

Chapter 6 was originally published with the title Translation Mining: Definiteness across languages. A reply to Jenks (2018) in Linguistic Inquiry, Vol. 53, and is a joint work with David Bremmers and Bert Le Bruyn. The original paper checked the empirical claims made by Jenks (2018). The paper plays the same role in the argumentation of this dissertation.

From indefinite contexts in Chapter 5, we turn to definite contexts in Chapter 6. As we pointed out in the conclusion of Chapter 3, our cross-linguistic corpus design has not allowed us to confirm or reject that the alternation we found between demonstratives and bare nouns in Chapter 4 is an artifact of our corpus. This is why, on the basis of the research overviewed in Chapter 4, we proposed two hypotheses: one that states there is an alternation and one that states there is no alternation. In Chapter 6, we investigate the hypothesis that there is an alternation and, more specifically, an alternation along the lines of the weak/strong definiteness distinction. This hypothesis was defended by Jenks (2018).

As we argue in Chapter 6, a translation corpus approach is particularly suited to check whether a distinction that has been argued to be relevant for one language is also relevant for another language. The weak/strong definiteness distinction was originally proposed to account for the meaning difference between German definite articles that contract with prepositions and those that do not. Schwarz (2009) argued that the former count as uniqueness (or weak) definites, while the latter count as familiarity (or strong) definites. Jenks (2018) hypothesizes that the same distinction is active in Mandarin, with bare nouns counting as weak definites and demonstratives counting as strong definites.

The corpus study we report in Chapter 6 uses the full HP corpus in its German translation to arrive at a balanced set of contexts in which German uses contracted ($n = 40$) and uncontracted ($n = 56$) definite articles and compares these contexts to their Mandarin counterparts. If Jenks is correct about Mandarin and Schwarz is correct about German, we expect contracted definites to be rendered as bare nouns in Mandarin and uncontracted definites to be rendered as demonstratives. However, according to the results, even though the German data are by and large in line with the distinction

Schwarz proposes, the Mandarin data are far from showing the same tendency. On the one hand, we found confirmation that weak definites require bare nouns—Mandarin bare nouns occurring as the default counterparts of German contracted definites—but we did not find confirmation that strong definites require demonstratives. Instead, we found that while demonstratives are counterparts of German uncontracted definites, the vast majority of German uncontracted definites have bare nouns as their counterparts.

Thus, we conclude that the weak/strong definiteness alternation hypothesis that we proposed based on the research reported in Chapter 4 should be rejected. Even though this does not mean there is no alternation, it does invite us to take a serious look at the alternative hypothesis we proposed, *viz.*, that there is no real alternation between bare nouns and demonstratives and that the demonstratives we found in definite contexts in Mandarin really do count as demonstratives. This is the hypothesis we test in Chapter 7.

Chapter 7: Analyzing the Competition between Bare nouns and Demonstratives

In Chapter 7, we investigate whether the Mandarin demonstratives we found as translations of NPs preceded by *the* count as actual demonstratives rather than relying on an article-like use of demonstratives. Thus, we must determine what counts as a “regular” demonstrative, independently of cross-linguistic comparison. This is because the cross-linguistic comparison we carried out in Chapter 3 did not yield conclusive results.

The strategy we deploy in Chapter 7 is to adopt a recent formal analysis of demonstratives and definites and to check whether it allows us to predict the division of labor we found between bare nouns and demonstratives in definite contexts. We restrict our attention to strong definite contexts, as the corpus study reported in Chapter 6 showed that these are the definite contexts that contain both bare nouns and demonstratives. The data we relied on consist of all strong definites in the HP corpus that occur as first anaphoric pick-ups of referents introduced by NPs preceded by numeral-*yi* ($n = 64$). Moreover, we relied on Ahn (2022)’s analysis of anaphoric demonstratives and definites. For this analysis, we argue that there are two relevant differences between demonstratives and definites:

- (i) Anaphoricity is presupposed for the definite and asserted for demonstratives.

- (ii) The definite requires its NP to refer to a singleton set.

In the study, we translate these differences into referential and discourse properties of demonstratives and definites that we can annotate in our corpus. We can also check whether these properties allow us to predict the actual division of labor between bare nouns and demonstratives in our corpus.

The results of our corpus study show that the differences between anaphoric demonstratives and definites from Ahn's analysis allow us to predict the division of labor between bare nouns and demonstratives in our corpus. Thus, we conclude that the demonstratives that we find as translations of NPs preceded by *the* count as demonstratives. This chapter further reflects on why Mandarin demonstratives seem to have a wider use than demonstratives in other languages and relates this to differences in language-specific discourse-level preferences and language-specific constraints on contextual restrictions.

At the end of Chapter 7, the empirical, methodological, and theoretical work in this dissertation allows us to draw the following conclusions:

- Mandarin demonstratives that occur as translations of NPs preceded by *the* count as regular demonstratives, not as definites.
- Mandarin demonstratives do not alternate with bare nouns in definite contexts.
- Differences between the use of demonstratives in Mandarin and their use in other languages point to differences in language-specific discourse-level preferences and language-specific constraints on contextual restrictions.

In Chapter 7, we provide type-shifting based versions of Ahn's analyses of demonstratives and definites that allow our Mandarin data to straightforwardly be derived in the Properties Approach based on the following assumptions:

- Mandarin has no definite article and, therefore, allows bare nouns to freely undergo the ι type-shift to occur as definite arguments.
- Mandarin demonstratives count as demonstratives and do not interact with the basic type-shifts assumed in the formal semantic literature on argument formation.

We are now ready to turn to the overarching conclusions of this dissertation.

Chapter 8: Conclusion

In Chapter 8, we bring together our main methodological, empirical, and theoretical conclusions; reflect on the strengths and limitations of our work; and propose possible avenues for future research. For our argumentation overview, we restrict ourselves to a summary of the main conclusions, which we present by providing short answers to our three guiding research questions.

1. Empirically, how does Mandarin encode definiteness and indefiniteness? Do we only find bare nouns or also numeral-*yi* and/or demonstratives in indefinite and definite contexts? If we find both, what is the division of labor?
2. Theoretically, what do our empirical findings mean for formal semantic approaches to argument formation? Do we need to develop a fundamentally different alternative to existing approaches like the Kinds and Properties Approaches, or can we account for the alternations and derive the division of labor within at least one of these approaches?
3. Methodologically, how can we ensure that a translation corpus approach provides us with reliable data that allow us to answer our empirical questions?

Empirically, the research we describe in this dissertation confirms that numeral-*yi* and bare nouns alternate in indefinite contexts (Chapters 2 and 3)—the former in regular argument position and the latter in pseudo-incorporated position (Chapter 5). Our research also confirms that there is no real alternation between bare nouns and demonstratives in definite contexts, bare nouns counting as the only real definites and demonstratives functioning as real demonstratives (Chapters 2, 3, 6, and 7). Our empirical results thus align with the functionalist literature for the alternation of bare nouns and numeral-*yi* in indefinite contexts but not for the alternation of bare nouns and demonstratives in definite contexts. Even though this dissertation focuses on Mandarin, our translation corpus approach also led us to examine a range of other languages, Russian and Hindi in particular. For Russian, we found no alternations between bare nouns, demonstratives, and numeral-one. For Hindi, we found an alternation between bare nouns and numeral-one but not between bare nouns and demonstratives.

Theoretically, we conclude that our empirical outcomes for Mandarin can be accounted for within the Properties Approach (Chapters 4 to 7). This means that the alternation challenge does not require us to develop a fundamentally different alternative to existing formal semantic approaches to argument formation. Our Mandarin

data follow if we assume that numeral-*yi* functions as an indefinite article and if we adopt the analysis of pseudo-incorporation we developed in Chapter 5 and the difference between demonstratives and definites we described in Chapter 7 based on Ahn (2022). Even though our theoretical claims bear on Mandarin, we argued that they can easily be integrated with theoretical claims that allow the Properties Approach to derive the data patterns in Russian and Hindi as well (Chapter 4). This argues in favor of the cross-linguistic robustness of the Properties Approach and its value as a general theory of argument formation. Our conclusion that the Properties Approach can account for the data in Mandarin, Russian, and Hindi does not imply that the Kinds Approach would be unable to do so but merely that evaluating whether two approaches can account for the data would have led us beyond the scope of this dissertation.

Methodologically, we conclude that we can ensure that a translation corpus approach provides reliable data if we consider a broad range of data from the source language and opt for a radically cross-linguistic setup and not for a contrastive one. By looking at a broad range of data from the source language, we made sure we investigated bare nouns, numeral-*yi*, and demonstratives in their own right, and we were able to confirm the target language representativeness of our Mandarin translation through benchmark categories (Chapter 2). By taking a radically cross-linguistic approach, we were able to look for unwanted biases in our corpus (Chapter 3) and argue for the stability of definiteness and indefiniteness across translations (Chapter 4). Taken together, these methodological results led us to conclude that a translation corpus approach, if modeled as described in this dissertation, is a viable method for investigating argument formation within and across languages.

1.6 Conclusion

This chapter began by setting up the **alternation challenge** as the central puzzle that this dissertation aims to resolve (Section 1.1). Functionalist observations report that Mandarin bare nouns alternate with demonstratives in definite contexts and with numeral-*yi* in indefinite contexts, whereas current formalist accounts of argument formation (the Kinds Approach and the Properties Approach) predict that bare nouns should be sufficient in regular (in)definite contexts. In Section 1.2, we outlined the theoretical background underlying this alternation challenge. Then, we formulated the guiding empirical and theoretical questions following this tension in Section 1.3 and motivated a translation corpus methodology in Section 1.4 to anchor definite and indefinite expressions in Mandarin via its translations of English *the* and *alan* as se-

mantic proxies of regular definite and indefinite contexts. Section 1.5 outlined the argumentation structures in each chapter outlined; we also demonstrated that the remainder of the dissertation proceeded to accomplish three goals: (i) to empirically establish the real Mandarin forms and where alternations actually occur in (in)definite contexts, (ii) to evaluate which formal analysis derives the observed patterns with the division of labor, and (iii) to ensure the reliability of the translation corpus approach in the referential domain. With these goals in place, we start from the empirical baseline in Chapter 2.

CHAPTER 2

A Corpus Investigation of the Mandarin Referential System

2.1 Introduction

Chapter 1 established the alternation challenge as the central puzzle of this thesis's exploration of Mandarin (in)definiteness. The functionalist literature claims that Mandarin bare nouns alternate with numeral-*yi* constructions in indefinite contexts and with demonstratives in definite contexts. However, this functionalist view contrasts with predictions of formal semantic frameworks (both the Kinds Approach and the Properties Approach) that perceive Mandarin bare nouns as self-sufficient arguments and predicts no alternation should occur in definite or indefinite contexts.

This chapter moves from claims made in the literature to language data and starts to tackle our overarching empirical question regarding how Mandarin encodes definiteness and indefiniteness and whether we only find bare nouns or if we also find numeral-*yi* and/or demonstratives in indefinite and definite contexts. However, before we address any theoretical implications of the alternation challenge, we must first confirm whether the alternations—namely, the co-occurrence of bare nouns with numeral-*yi* in indefinite contexts and with demonstratives in definite contexts—are an empirical reality and systematically occur in natural use of Mandarin. Without this empirical baseline confirming the reality and systematicity of the alternations in this

chapter, the theoretical debates explored in later chapters would remain unstable and tentative.

To achieve this, we employ the translation corpus methodology outlined in Section 1.4, using the first chapter of *Harry Potter and the Philosopher's Stone* and its Mandarin translation. Rather than pre-selecting for definite or indefinite contexts, we examine the whole Mandarin referential system corresponding to English counterparts. This approach allows us to map the distribution patterns of Mandarin expressions in their own right, not merely as translations of English NPs preceded by *the* and *alan* (subquestion 3.1). Within this broad investigation, the translation of English definite and indefinite articles serves as a crucial point of analysis. To ensure the validity of this method, we set up benchmark categories to assess whether our empirical results reflect established patterns in the literature about Mandarin (subquestion 3.3).

To guide our analysis, we derive a set of testable hypotheses from the literature. These hypotheses integrate claims from the functionalist perspectives on Mandarin forms encoding definiteness and indefiniteness, as well as those from the formalist literature on Mandarin argument formation. Specifically, we test the following five hypotheses:

- Hypothesis 1: Bare nouns occur in both definite and indefinite contexts (Chierchia, 1998; Dayal, 2004; Li and Thompson, 1989; Wright and Givón, 1987).
- Hypothesis 2: Indefinite bare nouns are restricted to the object position (Cheng and Sybesma, 1999).
- Hypothesis 3: Demonstratives occur in definite contexts and in demonstrative contexts (Chen, 2004).
- Hypothesis 4: Numeral-*yi* constructions occur in indefinite contexts and numeral contexts (Wright and Givón, 1987; Chen, 2003, 2004).
- Hypothesis 5: Bare nouns, demonstrative and numeral-*yi* constructions comprise the full and exhaustive set of expressions occurring in Mandarin definite and indefinite contexts (Li and Thompson, 1989).

This chapter is structured as follows. Section 2.2 details the translation corpus methodology and derives predictions from the above hypotheses. Section 2.3 presents the quantitative results, beginning with our benchmark categories before moving to the overview of English-Mandarin referential system alignments and then to the core data on definite and indefinite contexts. Section 2.4 discusses the implications of these

findings. Finally, Section 2.5 concludes the chapter by summarizing what this empirical baseline means for the alternation challenge and by setting up the next step of the investigation.

2.2 The translation corpus

This section details the translation corpus methodology used to test the five hypotheses outlined in Section 2.1. By systematically mapping the referential systems of English and Mandarin, we aim to gather the quantitative evidence necessary to empirically assess the alternation challenge. The following subsections will describe the procedures for corpus construction, the establishment of benchmark categories for validation, and the specific predictions derived from our hypotheses.

2.2.1 Corpus construction, data collection, and annotation

We draw data from Chapter 1 of *Harry Potter and the Philosopher's Stone* in its English original and its Mandarin translation. English and Mandarin sentences were aligned, and the accuracy of the sentence alignment was manually checked by the author.

From the aligned sentence pairs, we extracted English referential expressions in canonical argument positions (i.e., subjects, direct/indirect object positions, and complements of prepositions). The data extraction process includes (i) benchmark categories (i.e., English personal pronouns and possessive constructions); (ii) English singular definites (NPs preceded by *the*) and singular indefinites (NPs preceded by *a/an*); (iii) other major English referential categories. For each English expression, we identified and extracted its corresponding expressions in the aligned Mandarin translations. For details of the extracted data, see Appendix A.

Data annotation followed a form-driven procedure that was applied consistently across both English and Mandarin. All extracted English referential expressions and their Mandarin alignments in translations were annotated based on their forms. Each distinct English form was coded. Likewise, Mandarin aligned expressions in translations were coded by forms. These forms were then grouped into semantic categories for analytical convenience. For example, in English, *the*+N was classified as definite, *a/an*+N as indefinite, *his*+N as possessive, and so on. In Mandarin, forms were mapped to their referential categories. For instance, N was classified as a bare noun, *zhè/nà*+CL+N as demonstratives, *yì*+CL+N as numerals. Full details of the annotation

results of both English expressions and their Mandarin alignments are presented in the Appendix A.

2.2.2 Benchmark categories

To ensure that our translation corpus is reasonably representative of Mandarin, we first evaluate whether our corpus data can reliably capture the expected correspondences between English and Mandarin referential forms in the literature. We do this by analyzing two benchmark categories before turning to our focus to definite and indefinite expressions.

The two benchmark categories were chosen based on existing literature to represent stable and variable alignment patterns. By the stable correspondence categories, we mean English forms that are expected to map predominantly to a single primary Mandarin form. We selected English personal pronouns as the stable benchmark category because previous research suggests that English personal pronouns generally align with Mandarin personal pronouns in a one-to-one manner, despite language specific features such as pro-drop or minor non-canonical alignment (Li and Thompson, 1989; Xiang, 2019; Hsiao, 2011). Meanwhile, variable correspondence categories contain English expressions that are known to align with multiple Mandarin forms. We chose English possessives (including possessive determiners and 's possessive constructions) as the variable benchmark category since English possessives are known to align with multiple Mandarin forms, including possessive constructions (with the overt possessive marker *de*), bare nouns, and demonstratives (Li and Thompson, 1989; Partee, 2008; Niu, 2015).

We believe that if the corpus can successfully capture these benchmark referential categories, as reflected in the stable pattern of pronouns and the variable pattern of possessives, then the translation corpus can represent the target language, Mandarin. Thus, we could assume that the corpus provides a solid basis for capturing patterns in definite and indefinite domains.

2.2.3 Predictions

Based on the hypotheses outlined in Section 2.1, we develop five testable predictions on the alignment patterns we expect to find in the translation corpus:

- Prediction 1: Mandarin bare nouns will be found as translations for English expressions in both definite contexts (i.e., in line with English *the+N*) and in-

definite contexts (i.e., in line with English *a/an+N*).

- Prediction 2: Mandarin bare nouns occurring in indefinite contexts will be restricted to the object position.
- Prediction 3: Mandarin demonstratives will align with English *the+N* in definite contexts and with English *this/that+N* in demonstrative contexts.
- Prediction 4: Mandarin numeral-*yi* constructions will align with English *a/an+N* in indefinite contexts and align with English numeral-one in numeral contexts.
- Prediction 5: Mandarin bare nouns and demonstratives are the only forms that will occur in definite contexts aligned with English *the+N*, and Mandarin bare nouns and numeral-*yi* constructions will be exhaustive forms occurring in indefinite contexts aligned with English *a/an+N*.

These predictions are tested with the corpus results in terms of their quantitative distributional frequencies in Section 2.3.

2.3 Results

The section is organized as follows. First, Section 2.3.1 presents the results of benchmark categories. Section 2.3.2 then provides a general overview of the distributions of English referential expressions and their Mandarin translations from the corpus. Next, Section 2.3.3 presents the results to test each of the predictions. Finally, Section 2.3.4 summarizes the main quantitative observations from this chapter.

Sankey diagrams are used to visualize all alignment results between English and Mandarin referential systems. The analysis is monodirectional from English to Mandarin, with frequencies shown by the numbers and the width of the flows. The following subsection provide more details.

2.3.1 Benchmark categories

First, to check the representativeness of our translation corpus data, we examined the alignment patterns for the two benchmark categories: the stable correspondence (namely, English personal pronouns) and the variable correspondence (namely, English possessives, including possessive determiners and 's possessive constructions).

We first examine the alignment results for the stable correspondence category with personal pronouns using the Sankey diagram in Figure 2.1. As shown in the left panel

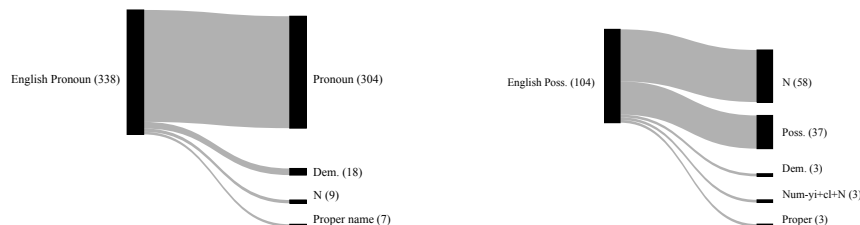


Figure 2.1: Alignments of benchmark categories with stable correspondence vs. variable correspondence.

of Figure 2.1, English personal pronouns ($n = 338$) overwhelmingly align with Mandarin personal pronouns ($n = 304$), confirming the expected stable correspondence. Minor variations, such as English personal pronouns aligning with a Mandarin bare noun ($n = 9$), a Mandarin demonstrative ($n = 18$), or a Mandarin proper name ($n = 7$), are infrequent and do not weaken the overall primary alignment within the personal pronoun category between English and Mandarin. Examples are shown in (1).

- (1) Benchmark 1: Stable correspondent category with personal pronouns.
- a. English personal pronouns → Mandarin personal pronouns (major alignment)

‘He supposed this was some stupid new fashion.’

tā cāixiǎng zhè dàgài yòu shì yì zhǒng wúliáo de xīn shíshàng
 he guess this probably again be one kind boring DE new fashion
 ba.
 PART
 - b. English personal pronouns → Mandarin bare noun (minor variation)

‘It was now sitting on his garden wall.’

zhè shí māo zhèng zuò zài tā jiā huāyuán de yuànqiáng
 this moment cat PROG sit at his home garden REL courtyard.wall
 shàng.
 on
 - c. English personal pronouns → Mandarin demonstrative (minor variation)

‘He was sure it was the same one; it had the same markings around its eyes.’

tā kěndìng zhè zhī māo hé zǎoshàng de nà zhī māo shì tóng yì zhī
 he sure this CL cat and morning REL that CL cat be same one

yǎnjīng zhōuwéi de wénlù yì mú yí yàng.
eye around REL marking exactly same

- d. English personal pronouns → Mandarin proper name (minor variation)
‘...but they’re saying that when **he** couldn’t kill Harry Potter, Voldemort’s power somehow broke—and that’s why he’s gone.’

búguò tāmen shuō dāng **Fúdimó** shā bù sǐ Hǎlì de shíhòu tā
but they say when Voldemort kill NEG die Harry POSS time he
de gōngfǎ jiù bùzhī zěn de shǐlíng le suǒyǐ tā
POSS magic.method then NEG-know how REL malfunction ASP so he
cái zǒudiào le.
only leave.away ASP

- e. English definites → Mandarin personal pronoun (a case for translator’s choice)

‘In fact, it was nearly midnight before **the cat** moved at all.’

shíjìshàng kuài dào wǔyè shí **tā** cái kāishǐ dòng le dòng.
actually nearly reach midnight moment it then begin move ASP move

Examples in (1) demonstrate how English personal pronouns align with expressions in Mandarin. Example (1a) shows the primary alignment pattern, with the English personal pronoun *he* directly corresponding to the Mandarin personal pronoun *tā* (‘he’). Examples (1b)–(1d) are cases for exceptional alignments: in (1b), the English personal pronoun *it* aligns to the Mandarin bare noun *māo* (‘cat’); in (1c), *it* corresponds to the Mandarin demonstrative *zhè zhī māo* (this CL cat); in (1d), *he* is rendered as the proper name *Fúdimó* (‘Voldemort’) in the Mandarin translation. These exceptional cases account for a small portion of our dataset compared to the primary alignment between English personal pronouns ($n = 338$) and Mandarin personal pronouns ($n = 304$). We argue that these cases primarily reflect the translator’s choice, a flexibility also evident when, conversely, other English forms are translated as Mandarin personal pronouns, as in (1e). Therefore, we confirm that our parallel translation corpus reliably captures stable correspondence categories in the referential system.

As for the other benchmark—the variable correspondent category—English possessive constructions ($n = 104$) align with multiple Mandarin forms, as shown in the right panel of Figure 2.1. The most frequent alignments from English possessives are to Mandarin bare nouns ($n = 58$) and Mandarin possessive constructions ($n = 37$). Minor variations include demonstratives ($n = 3$) and numeral-*yi* ($n = 3$). Examples in (2) illustrate some cases for these alignments in our dataset.

- (2) Benchmark 2: Variable correspondent category with possessives.

a. English possessive → Mandarin possessives

‘As he pulled into the driveway of number four, the first thing he saw—and it didn’t improve **his mood**—was the tabby cat he’d spotted that morning.’

dāng tā shǐrù sì hào chēdào shí,
when he drive.into four number lane when,

dì.yī gè yìng rù yǎnlián de jiù shì
first CL enter into sight REL just be

zǎoshàng tā jiàn guò de nà zhī huābānmāo,
morning he see PERF REL that CL tabby.cat,

zhè bìng méiyǒu shǐ tā de xīnqíng hǎozhuǎn.
this at.all not.have make he GEN mood improve

b. English possessives → Mandarin bare nouns

‘At half past eight, Mr. Dursley picked up **his briefcase**, pecked Mrs. Dursley on the cheek and tried to kiss Dudley goodbye but missed, because Dudley was now having a tantrum and throwing his cereal at the walls.’

bā diǎn bàn, Désiǐ xiānsheng ná qǐ gōngwénbāo,
eight o'clock half, Desli Mr. pick up briefcase,

zài Désiǐ tàitai miànjiá shàng qīn le yí xià,
at Desli Mrs. cheek on kiss PERF one time,

zhèng yào qīn Dáì, gēn zhè gè xiǎo jiāhuo dàobié,
just about.to kiss Dalì, with this CL little kid say.goodbye,

kěshì méiyǒu qīn chéng, xiǎo jiāhuo zhèngzài fā píqì,
but not kiss succeed, little kid PROG explode temper,

bǎ mài piàn wǎng qiáng shàng shuāi.
BA cereal toward wall on throw

c. English possessives → Mandarin demonstratives

‘In **his vast, muscular arms** he was holding a bundle of blankets.’

tā nà jīròu fādá de cūzhuàng shuāngbì bào zhe yì juǎn
he that muscle develop REL stout pair.arms hold DUR one CL
máotǎn.
blanket

d. English possessives → Mandarin numeral-*yi* constructions

‘His blue eyes were light, bright and sparkling behind half-moon spectacles ...’

bànyuèxíng de yǎnjìng hòubiān yí duì zhànlán zhànlán
 half.moon.shaped REL glasses behind one pair deep.blue deep.blue
de míngliàng yǎnjīng shǎnshǎn fàng guāng
 REL bright eyes twinkling emit light

In (2a), *his mood* aligns with the Mandarin possessive construction *tā de xīnqíng* (he DE mood), showing a direct possessive-to-possessive correspondence. In (2b), *his briefcase* aligns with the bare noun *gōngwénbāo*; in (2c), *his vast, muscular arms* aligns with *tā nà jīròu fādá de cūzhuàng shuāngbì* (he that muscle developed arms), where *his* is rendered by the pronoun *tā* and the demonstrative *nà*; in (2d), *his blue eyes* corresponds to a numeral-*yi* construction, *yí duì zhànlán zhànlán de míngliàng yǎnjīng* (one CL blue DE bright eyes). These varied alignments within our corpus reflect the known flexibility of how English possessives are translated into Mandarin possessives as well as other variations in multiple forms, which proves our corpus’s ability to capture variable correspondence patterns.

These benchmark results show that our parallel translation corpus and the alignment methodology can capture both stable and variable correspondent patterns as described in the literature. Having established the reliability of our corpus and methodology through these benchmark categories, we now overview the entire referential system to situate our primary objects of study—namely, definiteness and indefiniteness—within the general referential system. In the next subsection, we show the general alignment of the whole referential systems between the two languages to prepare for a broader landscape for definite and indefinite expressions.

2.3.2 Overall distribution of referential expressions

Figure 2.2 below shows the alignment for the general English referential system to the Mandarin system with a Sankey Diagram ($n = 1210$). Before we assess the direct correspondences between referential forms, we must first clarify the nature of this large “untranslated” category ($n = 216$), which constitutes a significant portion of the data.

This category mainly includes instances for which the English referential expression is either omitted in the Mandarin translation due to pro-drop or topic ellipsis, or it is rendered with an idiomatic expression in Mandarin that has no direct counterpart in

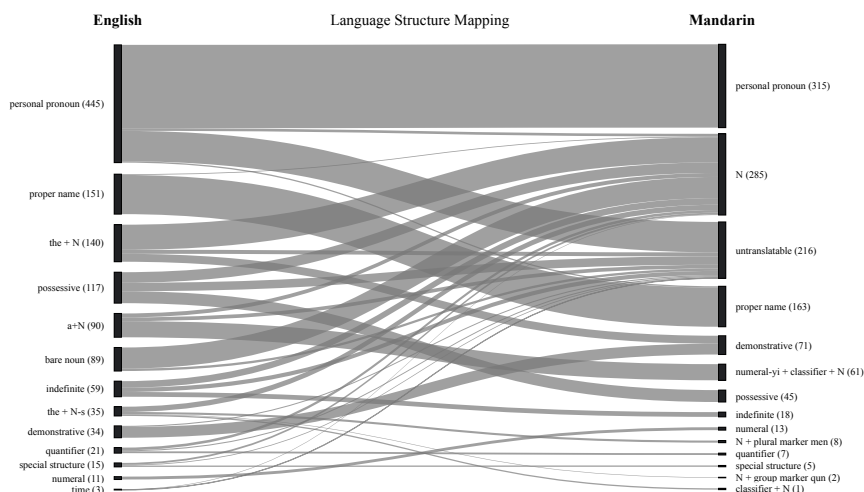


Figure 2.2: Alignments of English and Mandarin Referential Expressions.

the English original. While a full analysis of these zero forms or idiomatic expressions is beyond the scope of our present form-based study, they constitute a significant component of the Mandarin referential system and warrant future investigation. Examples for the untranslated cases are shown in (3).

(3) Untranslated category

- a. ‘The Potters, that’s right, that’s what I heard.’

Bōtè fūfù, búcuò, wǒ zhèngshì tīngshuō.
Potter couple, not.bad, I exactly hear-say

- b. ‘The nerve of him!’

zhēn bù zhī xiūchǐ!
really not know shame!

Crucially, this untranslated category accounts for the discrepancy in totals between the overview in Figure 2.2 and the detailed analyses that follow. Since these expressions lack a correspondence between English and Mandarin forms, they are excluded from the analysis of alignment patterns. Specifically, this category includes 14 of the initial 140 definite NPs and 21 of the initial 90 indefinite NPs. Consequently, the dataset for our primary analysis in Section 2.3.3 consists of the remaining 126 definite and 69 indefinite expressions that received a translatable Mandarin counterpart.

Now that we have clarified the scope of our analysis, we can observe the general patterns among the interpretable alignments. As shown in Figure 2.2, English singular definites (*the*+N) and singular indefinites (*a/an*+N) distribute across multiple Mandarin expressions in a hybrid pattern. Specifically, English *the*+N aligns most frequently with Mandarin bare nouns and demonstratives, while English *a/an*+N aligns most frequently with Mandarin *yi*+CL+N and bare nouns. All other alignments are too marginal to indicate any tendency out of our current corpus. We can therefore conclude that the key forms discussed in the literature are the primary expressions used in the definite and indefinite domains. Crucially, our broad analysis of the referential system reveals no other Mandarin forms that compete in terms of distributional frequency, suggesting this set of expressions, *viz.*, bare nouns and numeral-*yi* in the indefinite contexts, and bare nouns and demonstratives in the definite contexts, is exhaustive for these specific contexts.

Having established that these forms are the primary expressions in the definite and indefinite domains, the following section describes the tests for each of the above specific predictions outlined in Section 2.2.3 regarding their distributional patterns.

2.3.3 Testing predictions: alignment in definite and indefinite contexts

Prediction 1: Bare nouns occur in definite and indefinite contexts

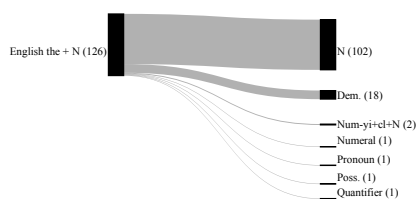
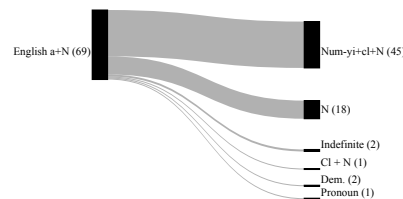
Prediction 1 states that Mandarin bare nouns will occur in both definite and indefinite contexts; it also predicts that Mandarin bare nouns should align with both English singular definites and indefinites.

Our corpus results support this prediction. As shown in Figure 2.3, Mandarin bare nouns align with English *the*+N in 102 out of 126 cases. Furthermore, Figure 2.4 shows that Mandarin bare nouns align with English *a/an*+N in 18 of 69 cases. Examples for the alignment from English definites and indefinites to Mandarin bare nouns are shown in (4) and (5).

- (4) English definites → Mandarin bare noun

‘It was now reading **the sign** that said Privet Drive—no, looking at the sign.’

māo zhèshí zhèngzài dú nǚzhēn-lù de **biāopái**, bù, shì zài kàn
 cat now PROG read Privet-Drive DE sign, no, be PROG look
 biāopái.
 sign.

Figure 2.3: Alignments of English Definites (*the+N*) to Mandarin Expressions.Figure 2.4: Alignments of English Indefinites (*a/an+N*) to Mandarin Expressions.

(5) English indefinites → Mandarin bare noun

‘It was on the corner of the street that he noticed the first sign of something peculiar—a cat reading **a map**.’

zài jiē jiǎo shàng, tā kàn.dào le dì yī gè yìcháng de xìn hào ——
 on street corner on, he see.RVC LE the first CL peculiar DE sign ——
 yì zhī māo zài kàn dìtú.
 one CL cat PROG read map.

In (4), the bare noun *biāopái* (sign) is used to refer to *the sign*, which is unique and definite as known in the context. In (5), the Mandarin bare noun *dìtú* (map) is used as the correspondent translation to the English singular indefinite *a map*.

These alignment patterns provide clear empirical support for Prediction 1, as they show that bare nouns occur as Mandarin correspondent expressions to both English definites and indefinites. This finding aligns with the formalist and functionalist understanding of Mandarin bare nouns. The implications of this result are described in the discussion section of this chapter.

Prediction 2: Restriction of bare nouns in indefinite contexts

Prediction 2 anticipates a positional restriction on Mandarin bare nouns in indefinite contexts, limited to object positions (Cheng and Sybesma, 1999). Our data shows a lower frequency of bare nouns’ distribution in the indefinite contexts than in definite ones. We further annotated the syntactic position (subject vs. object) for Mandarin bare nouns aligned with English definites and indefinites. The results confirm the subject-object asymmetry (Table 2.1).

Table 2.1 shows that, if the data is compared only in the subject and object positions, bare nouns tend to be more frequent in definite contexts ($n = 60$) than in

Table 2.1: Mandarin bare noun Distribution across Subject/Object Positions in Definite/Indefinite Contexts.

Contexts	Subject position	Object position	Total
Definite context	14 (23.3%)	46 (76.7%)	60
Indefinite context	1 (9.1%)	10 (90.9%)	11
Total (subject/object)	15	56	–

indefinite contexts ($n = 11$). Crucially, in indefinite contexts, bare nouns appear in the subject position only once, compared to 10 times in the object position. Although with a general tendency of more bare nouns in object positions in both definite and indefinite contexts—and more definite data over indefinite ones in both English origins and accordingly Mandarin translations—the imbalance of bare nouns occurring in object position over subject position is particularly strong in indefinite contexts (1 subject vs 10 object). We show the one case for Mandarin bare nouns in the subject position aligned with an English NP preceded by *a* as in (6).

(6) Indefinite bare noun in the subject position

‘A breeze ruffled the neat hedges of Privet Drive, which lay silent and tidy under the inky sky, the very last place you would expect astonishing things to happen.’

wēifēng fú dòng zhe nǚzhēn-lù liǎng páng zhěngjié de shù lí...
breeze brush PROG Privet-Drive flank neat DE hedge...

In (6), *wēifēng* (‘breeze’) in the Mandarin sentence can be weak definites, as it refers to a kind individual while lacking a unique referent to a concrete object. Such cases of bare nouns occurring in the subject position are not surprising given its weak definiteness.

Prediction 3: Demonstratives in definite contexts and in demonstrative contexts

Prediction 3 posits that in definite contexts, in light of English *the*+N, Mandarin demonstratives occur as correspondents next to Mandarin bare nouns. This indicates bare nouns and demonstratives occur in definite contexts, and they also occur in demonstrative contexts aligned to English *this/that*+N, which indicates demonstratives still function with their canonical usages.

Our data show that both bare nouns and demonstratives are employed in definite contexts. Figure 2.3 shows the distribution of Mandarin expressions that correspond to

English *the+N* ($n = 126$). In line with the overview’s findings, bare nouns are the most frequent translation, accounting for 81% of cases ($n = 102$), while demonstratives are the second most common alternative, appearing in 14% of cases ($n = 18$). This confirms that in definite contexts, both forms are used. As shown in example (7), the English singular definite *the cat* is translated as the demonstrative *nà zhī māo* (that CL cat) in Mandarin.

(7) English definites → Mandarin demonstratives

‘As Mr Dursley drove around the corner and up the road, he watched **the cat** in his mirror.’

dāng dé.sǐ.lǐ xiān.shēng guǎi guò jiē.jiǎo jì.xù shàng.lù de
 when Dursley Mr. drive past street.corner continue on.road DE
 shí.hòu, tā cóng hòu.shì.jìng lǐ kàn.kàn **nà zhī māo**.
 time, he from back.look.mirror in watch that CL cat.

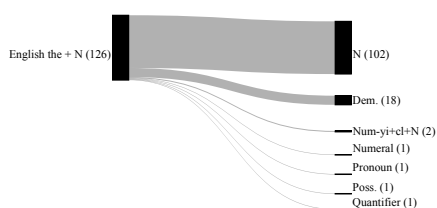


Figure 2.5: Alignments of English Definites (*the+N*) to Mandarin expressions (repeated from Figure 2.3).

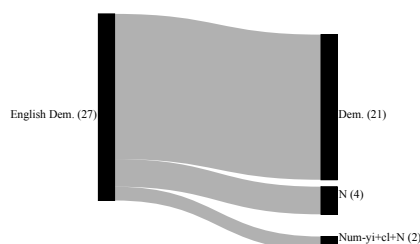


Figure 2.6: Alignments of English demonstratives to Mandarin expressions.

We also examined the alignment between English demonstratives (*this/that+N*) and Mandarin expressions. Comparing Figure 2.6 with Figure 2.5, we observe that while Mandarin demonstratives appear as expressions corresponding to English definites (Figure 2.5), they also primarily align to English demonstratives (Figure 2.6), in which Mandarin demonstratives correspond to English demonstratives in 21 out of 27 cases. This indicates that Mandarin demonstratives keep their canonical demonstrative function while extending to align with English definites in our corpus. Example (8) illustrates a Mandarin demonstrative aligning with an English demonstrative, in which English *that man* aligns to Mandarin *nàge nán de* (that CL man DE).

(8) English demonstratives → Mandarin demonstratives

‘Mr Dursley was enraged to see that a couple of them weren’t young at all;

why, **that man** had to be older than he was, and wearing an emerald-green cloak!’

...**nà gè nán de** xiǎnde bǐ tā niánlíng hái dà, jìngrán hái pī zhe
...that CL man DE appear than him age even big, actually still wear DUR
yí jiàn fěicù lǜ de dǒupéng!
one CL emerald green SUB cloak!

In the definite contexts, we also found exceptional cases where English definites aligned to Mandarin expressions in 2 cases were unexpected; an instance is shown in (9).

- (9) English definites → Mandarin numeral-*yi* construction (exceptional)
‘**The nearest street lamp** went out with a little pop.’

...lí de zuì jìn de **yì zhǎn lùdēng** pū de yì shēng
...away REL most near REL one CL streetlight puff REL one sound
xīmiè le.
extinguish PFV.

In (9), the English indefinite superlative *the nearest street lamp* aligns with a Mandarin numeral-*yi* construction. This can be attributed to the indeterminate semantics of the relative superlatives, which may not denote a uniquely identifiable individual in the same way as the absolute superlative (Coppock and Beaver, 2014) while allows for an existential interpretation expressed by the indefinite numeral-*yi* construction in Mandarin.

In summary, our quantitative findings demonstrate that Mandarin demonstratives serve a dual function: while they consistently fulfill their canonical demonstrative roles when aligned with English *this/that+N*, they also extend into the domain of definiteness, occurring alongside bare nouns as translations for English *the+N*. However, the precise extent to which Mandarin demonstratives have extended to function as definite markers—rather than simply retaining their deictic origins—remains an open question.

Prediction 4: Numeral-*yi* construction in indefinite contexts and in numeral contexts

Prediction 4 assumes that in indefinite contexts, Mandarin numeral-*yi* constructions will appear as expressions for English *a/an+N* alongside bare nouns. Meanwhile, numeral-*yi* constructions will also occur as canonical numerals aligned with English numeral-one. Our results confirm this prediction.

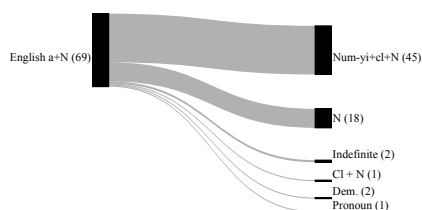


Figure 2.7: Alignments of English Indefinites (*a/an + N*) to Mandarin expressions (repeated from Figure 2.4).

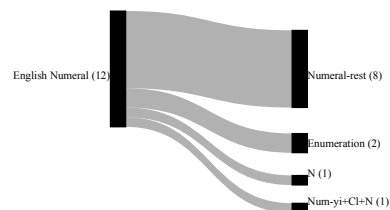


Figure 2.8: Alignments of English numerals to Mandarin expressions.

As shown in Figure 2.7, among the translations for English *alan+N* ($n = 69$), the Mandarin numeral-*yi* construction (*yi+CL+N*) is the dominant form, occurring in 65% of cases ($n = 45$). Bare nouns are the second most frequent translation, accounting for 26% of cases ($n = 18$). This shows that both forms function as indefinite expressions in our corpus. Example (10) presents a case in which the English indefinite (*a golden watch*) is translated with a Mandarin numeral-*yi* construction (*yí kuài jīnbǎo*).

(10) English indefinites → Mandarin numeral-*yi* construction

‘Dumbledore gave a great sniff as he took **a golden watch** from his pocket and examined it.’

Dèngbùliduō shēnshēn xī le yì kǒu qì, cóng yí dài lǐ tāo chū yì
 Dumbledore deeply breath.in LE one CL air from pocket in take out one
kuài jīn biǎo, rènzhēn kàn qǐ lái.
 CL golden watch seriously examine onward.

To further understand the role of the numeral-*yi* construction, we also examined the alignment between English numeral-one and Mandarin expressions, as shown in Figure 2.8 and example (11) below.

(11) English numeral-one → Mandarin *yi*

‘**One small hand** closed on the letter beside him and he slept on, ...’

tā de yì zhī xiǎo shǒu zhèng hǎo fāng zài nà fēng xìn pángbiān.
 he POSS one CL small hand just.right place at that CL letter beside.

There are also exceptional cases in the alignment between English indefinites and Mandarin expressions. An example of this kind of case is provided below, in which English indefinites align with Mandarin demonstratives. We argued that they are a

result of the translator's choice, and as such, they do not ruin the primary alignment tendency.

(12) English indefinites to Mandarin demonstratives (exceptional)

'Albus Dumbledore didn't seem to realise that he had just arrived in **a street** where everything from his name to his boots was unwelcome.'

Ābùsī Dèngbùlǐduō sīhū bìng méiyǒu yìshí dào cóng tā de míngzi
Albus Dumbledore seem indeed not.have awareness that from he DE name
dào tā de xuēzi, zài tā lái dào de **zhè tiáo jiē** shàng dōu bù shòu
to he DE boots, at he arrive DE this CL street on all not receive
huānyíng.
welcome.

The alignment of the English indefinite 'a street' to the Mandarin demonstrative *zhè tiáo jiē* (this CL street) in (12) reflects a translator's choice rather than a true outlier. The Mandarin demonstrative *zhè tiáo jiē* functions as a deictic demonstrative to emphasize the specific, immediate street Dumbledore has just arrived on, though this is not emphasized in the English original.

Crucially, comparing Figure 2.8 with Figure 2.7, we notice a much higher frequency of numeral-*yi* constructions as correspondent expressions for English singular indefinites than for English numeral-one. This indicates the significant role of the Mandarin numeral-*yi* construction in expressing indefinite meanings, in turn, confirms the complete grammaticalization of the numeral-*yi* construction as an indefinite marker beyond its numeral usage.

In parallel to the pattern seen with demonstratives, our results reveal that the Mandarin numeral-*yi* construction has moved substantially beyond its original cardinal function. While numeral-*yi* constructions occasionally correspond to the English numeral-one, they appear far more frequently as indefinite markers—aligning with English *a/an+N*—than in strictly numeral contexts. This indicates that the numeral-*yi* construction's primary function in the corpus as an indefinite marker, not a numeral. This distributional pattern contrasts with that of demonstratives, which, despite their expanded use in definite contexts, still align most frequently with their canonical English counterparts. We will return to this finding in the discussion section.

Prediction 5: The exhaustivity of bare nouns, demonstratives, and numeral-*yi* constructions

Prediction 5 posits that our target forms—bare nouns, demonstratives, and numeral-*yi* constructions—are the exhaustive expressions as correspondents to English singular

definites and indefinites. The results in Figure 2.3 and Figure 2.4 strongly support this prediction.

In definite contexts, as shown in Figure 2.3, Mandarin bare nouns ($n = 102$) and demonstratives ($n = 18$) together account for 120 out of 126 alignments for English *the+N* (95.2%). The remaining cases are distributed very rarely across five other categories (numeral, numeral-*yi*, personal pronoun, possessive, quantifier), with none of them representing a significant alternative.

Similarly, in indefinite contexts as shown in Figure 2.4, Mandarin bare nouns ($n = 18$) and numeral-*yi* constructions ($n = 45$) together account for 63 out of 69 alignments for English *alan+N* (91.3%). The other categories (CL+N, demonstrative, indefinite, personal pronoun) are again too rare to indicate other significant alternatives as primary indefinite expressions in our corpus.

In brief, our results strongly support Prediction 5: while a few marginal alignments exist, they do not challenge the overall finding; rather, they are specific translation choices or usages. Combined with the full landscape view of the general referential system between English and Mandarin outlined in Section 2.3.2, we can rule out the possibility of any other Mandarin forms that could compete in distributional frequency with our target expressions—bare nouns, demonstratives, and numeral-*yi* constructions—in either the definite or indefinite domain.

2.3.4 Recap of observations

This section presented a quantitative analysis of Mandarin referential forms aligned with English definite and indefinite articles. We first showed that benchmark categories are aligned with their Mandarin counterparts in patterns as predicted by existing literature, confirming the reliability of the methodology. Regarding the definite and indefinite contexts, Mandarin bare nouns were observed to occur as translations in definite contexts (aligned with English *the+N*) and indefinite contexts (aligned with English *alan+N*), supporting Prediction 1. Furthermore, Mandarin bare nouns were predominantly restricted to object positions in indefinite contexts, which supports Prediction 2. Moreover, Mandarin demonstratives occur in both definite and demonstrative contexts, supporting Prediction 3. We also found that Mandarin numeral-*yi* constructions occur in both indefinite and numeral-one contexts, supporting Prediction 4. Finally, the data confirm that bare nouns, demonstratives, and numeral-*yi* constructions form an exhaustive set of primary expressions for definite and indefinite meanings, supporting Prediction 5. Collectively, these findings present a hybrid pattern of

not only bare nouns, but also alternations as demonstratives and numeral-*yi* constructions co-existing in the Mandarin referential system aligned to “regular” definite and indefinite contexts in English originals. We explore the implications of these results in the discussion section.

2.4 Discussion

This chapter has established a quantitative empirical baseline for the Mandarin referential system in (in)definite contexts. By applying the translation corpus methodology outlined in Chapter 1, our results provide clear evidence for a systematic co-existence of forms: in definite contexts, bare nouns co-exist with demonstratives; in indefinite contexts, bare nouns co-exist with numeral-*yi*. This pattern empirically confirms the core observation from the functionalist literature, confirming that the alternation challenge is a real phenomenon.

In this section, by revisiting the five hypotheses outlined in Section 2.1, we demonstrate how our results distill the alternation challenge into detailed puzzles that motivate the remainder of this thesis.

First, the data confirm that bare nouns occur in both definite and indefinite contexts (Hypothesis 1) and that bare nouns along with demonstratives, and numeral-*yi* constructions, form the exhaustive set of primary expressions in these domains (Hypothesis 5). Our translation corpus findings thus move beyond claims in literature to provide quantitative evidence that Mandarin employs a hybrid system in encoding (in)definiteness. The high frequency of numeral-*yi* ($n = 45$) compared to low frequency of bare nouns ($n = 18$) in indefinite contexts ($n = 69$), as well as bare nouns ($n = 102$) compared to demonstratives ($n = 18$) in definite contexts ($n = 126$), confirms that the alternations described in the functionalist literature are core features of our data in both definite and indefinite contexts.

However, the results with raw frequencies raise a methodological challenge identified in Chapter 1 as how we interpret these numbers. The mere co-existence of these forms does not, by itself, confirm any principle governing the division of labor in the (in)definite domains. This prompts a further investigation into the principles governing their division of labor, while also raising the question of whether the observed pattern is merely an artifact of (random) context selection in the English original.

The confirmation of the remaining hypotheses allows us to sharpen this question by identifying two puzzles in indefinite and definite contexts. The confirmation of Hypothesis 2 reveals a critical distributional feature: bare nouns are severely constrained

in indefinite contexts, where they are predominantly licensed in object position; in comparison, such restriction is not observed in the definite domain where bare nouns take majority of definite contexts. The first puzzle emerges from the indefinite domain. The data offer particularly strong confirmation for Hypothesis 4, revealing a clear functional shift of the numeral-*yi* construction. Its high frequency as a translation for English NPs preceded by *a/lan* ($n = 45$) compared to its rare alignment with the numeral-one indicates a high degree of functional specialization as an indefinite marker in this corpus. Taken together with the distributional restriction of bare nouns in indefinite contexts as confirmed by Hypothesis 2, these two findings constitute the indefinite puzzle: the Mandarin indefinite domain is characterized by the co-existence of a highly frequent, functionally specialized indefinite marker, *viz.*, numeral-*yi* construction, and bare nouns whose use in indefinite contexts is restricted to the object position. This raises the central question to be addressed in Chapter 5: what principle governs the specific division of labor?

The second puzzle arises in the definite domain. Our data support Hypothesis 3 by showing that demonstratives serve a dual function. While we observed that demonstratives still take the majority of their canonical deictic roles ($n = 21$) in aligned English demonstrative contexts ($n = 27$), it contains 18 cases in the alignment to English NP preceded by the ($n = 126$). This pattern gives rise to the definite puzzle: unlike the numeral-*yi* construction, the demonstrative does not show a distinct shift in its function to be a definite marker; rather, it retains more in demonstrative contexts while coexists with far more frequent bare nouns in definite contexts. This leads to a guiding question for Chapters 6 and 7: what precise factors govern the division of labor between bare nouns and demonstratives in definite contexts, while retaining demonstratives still deictic, demonstrative?

In sum, by confirming the five initial hypotheses, this chapter has not only validated the functionalist observation of alternation but has also refined it into two specific, empirically-grounded puzzles. Crucially, this process has exposed the inherent limits of a bilingual corpus study. We have identified patterns, but we lack the criteria to interpret them. We do not know whether the distributional percentages in this corpus are significant enough to be considered genuine referential features. Instead, they may be an artifact of our selection of contexts in the English original. As outlined in Section 1.4, our solution is to involve the comparative corpus work with patterns in other languages. This is precisely the task of the next two chapters. In addition to checking for corpus bias, they represent the next step required to establish the cross-linguistic baseline against which the Mandarin data can be properly evaluated. Only after this

comparative work is completed can we determine if these puzzles in Mandarin are language-specific phenomena that demand the theoretical explanations developed in the later chapters of this thesis.

2.5 Conclusion

In this chapter, we conducted a quantitative analysis with the initial translation corpus study outlined in our methodological plan (Section 1.4), investigating our guiding research question (Section 1.3): how does Mandarin encode definiteness and indefiniteness? Mapping the English referential system to its Mandarin translations in the first chapter of *Harry Potter and the Philosopher's Stone* enabled us to determine which expressions Mandarin translations chose to render referential categories like definiteness, indefiniteness, demonstratives, and numerals, etc., in the English original. The patterns found in the translation corpus show that the Mandarin pattern does not consist only of bare nouns but that it contains Mandarin demonstratives and numeral-*yi* that extend beyond their canonical demonstrative and numeral contexts and play a role in definite and indefinite domains, respectively. This result provides the first quantitative, corpus-grounded validation of the core observation of the functionalist observation, confirming that the alternation challenge is a real phenomenon that any formal theory must address.

Methodologically, the design of the study described in this chapter addresses two of our guiding methodological questions. By analyzing the entire referential system rather than pre-selecting only definite and indefinite contexts, we ensured that we did justice to Mandarin expressions in their own right, thus providing a direct answer to Subquestion 3.1. Furthermore, the fact that our benchmark categories perform as predicted in the literature provides independent evidence that our translation data represent the target language, thereby answering Subquestion 3.3.

However, while our data support the existence of these alternations and answer the above initial methodological subquestions, our findings bring an interpretative challenge to the front, as anticipated by our methodological Subquestions 3.2 and 3.4 in Section 1.4. Specifically, we cannot tell if the observed alternation between bare nouns and demonstratives/numeral-*yi* reflect a genuine feature of the Mandarin referential system or if it reflects a corpus bias induced by our random selection of the English source text. Thus, we need to find a way to ensure that potential meaning differences between originals and translations do not undermine the validity of conclusions based on translated data.

As prefigured in our methodology in Section 1.4, addressing this challenge requires the multilingual, comparative approach involving other languages in translation corpora using the same English corpus based on *Harry Potter and the Philosopher's Stone* and its translations. This comparative approach is carried out in the next two chapters. The two published papers reproduced in Chapters 3 and 4 bring in cross-linguistic translation corpora to test whether the Mandarin pattern we found in this chapter is reliable and unique.

Chapter 3 begins this validation process by applying the same translation corpus methodology to Russian and Hindi, which, like Mandarin, are considered articleless languages that rely on bare nouns expressing definiteness and indefiniteness (Dayal, 2004). We use the same English corpus, *viz.*, English NPs preceded by *the* and *a* in the first chapter of *Harry Potter and the Philosopher's Stone*, as proxies for regular definite and indefinite contexts, and for mapping the distribution forms in Russian and Hindi translations. By comparing how bare nouns and demonstratives/numeral-one in these typologically similar languages translate the same English source contexts, we can isolate language-specific patterns from potential corpus-biased artifacts. The central question is whether they replicate the specific Mandarin alternation pattern, *i.e.*, the co-existence of bare nouns with demonstratives in definite contexts and with numeral-*yi* in indefinite contexts with close distributional percentages. If Russian and Hindi show distinct patterns, it would provide strong evidence for the reality of the alternation challenge as a Mandarin specific phenomenon. It would also confirm that there is no corpus bias at play.

Chapter 4 completes the validation by addressing methodological Subquestion 3.4 concerning how meaning stability can be ensured across translations. By expanding the comparison to include languages with full article systems (Spanish, German) and a language with only a definite article (Hebrew), we can test whether translation patterns in these languages align with their known grammatical rules for definiteness and indefiniteness. This provides strong grounds for assuming that definiteness and indefiniteness are stable in translation corpora. Through this broader typological lens, we can empirically re-test the patterns identified in Chapters 2 and 3. This step is essential for establishing the reliability of our corpus data and for theoretically evaluating how the major formal semantic frameworks—namely, the Kinds Approach and the Properties Approach—account for the cross-linguistic patterns observed in Mandarin and other languages.

Only after the rigorous empirical and theoretical groundwork in Chapters 3 and 4 can we develop the detailed formal analyses for Mandarin that follow in Chapters 5

to 7. Chapter 5 develops a theoretical analysis of the division of labor between Mandarin bare nouns and numeral-*yi*, while Chapters 6 and 7 examine the division of labor between Mandarin bare nouns and demonstratives.

CHAPTER 3

“Articleless” Languages are Not Created Equal¹

3.1 Introduction

In his seminal paper on reference to kinds across languages, Chierchia (1998) defends the intuition that languages that do not have articles freely allow their bare nouns (BNs) to give rise to definite and indefinite interpretations. On the basis of fine-grained data from Hindi, Russian and Mandarin, Dayal (2004) argues that this generalization holds for bare plurals (BPs) and for BNs in classifier languages, but not for bare singulars (BSs), the latter being restricted to definite interpretations. Dayal accounts for this new generalization in an updated version of Chierchia’s neo-Carlsonian framework.

In this chapter, we retrace the data underlying Dayal’s argumentation and sketch the way she accounts for them (Section 3.2). Following up on problematic data from Russian, we propose a translation corpus study based on the first chapter of *Harry Potter and the Philosopher’s Stone* and its translations to the three languages studied by Dayal (2004) (Section 3.3). Our Hindi data turn out to be overall in line with Dayal’s predictions but the same does not hold for our Russian and Mandarin data (Section

¹Chapter 3 was originally published with the title “Articleless” languages are not created equal in the proceedings of *Sinn und Bedeutung 27* and is joint work with Shravani Patil, Daria Seres, Olga Borik and Bert Le Bruyn. The original paper is positioned as an evaluation of Dayal (2004). The precise role of this chapter within the overall argumentation is outlined in Section 1.5.

3.4), leading us to explore a number of extensions and modifications of Dayal’s analysis (Section 3.5 and Section 3.6). The overall conclusion we will arrive at is that BNs do not behave in the same way across so-called “articleless” languages and that the explanation might lie in the fact that some of these languages are less articleless than the literature has suggested up till now.

3.2 BNs in “articleless” languages: from Hindi to Russian and Mandarin

Dayal (2004) argues that Hindi BNs display a singular/plural asymmetry. Whereas plural BNs (BPs) straightforwardly allow for narrow scope indefinite readings, singular BNs (BSs) turn out to be more restricted. We illustrate this asymmetry with the minimal pair in (1) (see Dayal, 2004):

- (1) a. #*caroN.taraf cuuha* hai
 everywhere mouse is
 ‘The same mouse was everywhere.’
- b. *caroN.taraf cuuhe* haiN
 everywhere mice are
 ‘There were mice everywhere.’

The intended readings of *cuuha* in (1a) and *cuuhe* in (1b) are those in which they take scope under *caroN taraf*, leading to the assertion that there were mice everywhere. As Dayal points out, this reading is available for (1b) but not for (1a), the latter only leading to the pragmatically odd assertion that the same mouse was everywhere. With Dayal, we conclude that the opposition between (1a) and (1b) shows that Hindi BSs do not have the same range of indefinite readings as Hindi BPs.

Even though the contrast in (1) seems to bear on scope in the indefinite domain, Dayal (2004) gives it a definiteness twist, arguing that the data follow if we assume BPs allow for indefinite interpretations but BSs do not. In Section 3.2.2, we develop this intuition in more detail, sketch how Dayal derives it in an extended version of the neo-Carlsonian framework and explore the implications for Hindi, Russian and Mandarin. To properly frame the extensions Dayal proposes, we however start by taking another look at the original version of the neo-Carlsonian framework (Chierchia, 1998) and the predictions it makes about Hindi BSs and BPs.

3.2.1 Chierchia’s predictions for Hindi BSs and BPs

Chierchia (1998) does not explicitly treat Hindi, but we can generate the predictions he makes for its BSs and BPs. For presentation purposes, we work out the predictions under Dayal’s assumption that Hindi is an articleless language, an assumption that Chierchia (1998) does not commit to.

For Hindi BSs, Chierchia’s neo-Carlsonian framework presents two ways to derive indefinite readings. The first is a simple existential shift (\exists): under Dayal’s assumption that Hindi is an articleless language, BSs are predicted to be able to undergo a covert \exists -shift and end up with an indefinite interpretation. The second way is a more involved one, building on Chierchia’s Derived Kind Predication (DKP) and the fact that Hindi BSs can refer to kinds, as illustrated in (2) (see Dayal, 2004, 402):

- (2) **kutta** aam janvar hai
 dog common animal is
 ‘The dog is a common animal.’

DKP is an operation that kicks in when a kind combines with a predicate requiring reference to regular individuals (see Chierchia, 1998, 364):

- (3) Derived Kind Predication (DKP):
 If P applies to regular individuals and k denotes a kind, then
 $P(k) = \exists x[\cup k(x) \wedge P(x)]$

On its definition in (3), DKP leads to existential quantification over instantiations of a kind, effectively giving rise to a (derived) indefinite reading. With BSs referring to kinds in Hindi, Chierchia’s DKP thus constitutes a second path to indefinite readings for BSs.

Moving to Hindi BPs, Chierchia predicts them to give rise to indefinite readings on a par with Hindi BSs. The one difference with BSs resides in the fact that the latter have two paths that lead to indefinite readings – the existential and the DKP path – whereas BPs only have the DKP path at their disposal. The full derivation of the DKP path starts with a shift from predicates to their corresponding kinds (the down-shift, \cap) and is followed by DKP. The reason BPs do not have the existential path at their disposal is that Chierchia ranks the shift from predicates to their corresponding kinds (the down-shift, \cap) above the \exists - and the iota (ι)-shifts, and argues that the \cap -shift is defined for plurals but not for singulars. Given that the \cap -shift is defined for BPs, its ranking above the \exists -shift blocks the latter from applying and cuts off the existential

route to indefinite readings for BPs. For BSs, the \sqcap -shift is undefined, and its higher ranking has no effect on the availability of \exists -shift, maintaining the latter as a path towards indefinite readings.

Summarizing Chierchia’s predictions for Hindi, we have worked out how BSs can get indefinite interpretations through the \exists -shift and DKP whereas BPs get indefinite readings through DKP alone. Importantly, though, the opposition in the availability of the \exists -shift has no bearing on the asymmetry we find in (1). Indeed, both the \exists -shift and DKP are expected to allow for narrow scope readings, leaving the unavailability of the narrow scope reading of the BS in (1a) and its asymmetry with the BP in (1b) unaccounted for.

3.2.2 Dayal’s account and predictions

Dayal’s extensions of Chierchia’s neo-Carlsonian framework are mainly targeting the singular paradigm. We present the underlying intuition and discuss the extensions Dayal proposes, focusing on BSs but also briefly looking into BPs. Dayal’s account is inspired by the intuition that Hindi BSs cannot get indefinite readings but only definite ones, straightforwardly explaining why *cuuha* in (1a) cannot but refer to a unique mouse and lead to the pragmatically odd assertion that the same mouse was everywhere. To derive this restriction to definite readings for BSs, Dayal introduces two extensions to Chierchia’s neo-Carlsonian framework. The first is to not only rank the \sqcap -shift above the \exists -shift but to do the same for the ι -shift, leading to the ranking $\sqcap, \iota > \exists$. The effect of this move is that the \exists -shift no longer constitutes a viable path to indefinite readings for Hindi BSs—independently of the fact that the \sqcap -shift is not defined for them. The second extension Dayal proposes is to restrict the availability of DKP to kinds that have a “semantically transparent relation to their instantiations” (Dayal, 2004, 430), a property that Dayal associates with kind reference of plural nouns but not of singular nouns. The effect of this second extension is that DKP is also cut off as a viable path to indefinite readings for Hindi BSs.

With the two extensions she proposes, Dayal makes sure that there are no paths to indefinite readings for Hindi BSs in her updated version of Chierchia’s neo-Carlsonian framework. She thus guarantees that the only non-kind referring readings BSs can get in regular argument position are definite ones, deriving the pragmatically odd reading of *cuuha* in (1a). For BPs, Dayal’s extensions have no impact on the availability of DKP-generated indefinite readings. The narrow scope indefinite reading Chierchia predicts for *cuuhe* in (1b) is thus maintained and the contrast with (1a) accounted for.

Dayal’s extensions of Chierchia’s neo-Carlsonian framework make a number of predictions. First, for Hindi BSs, the prediction is that they should never give rise to indefinite readings in regular argument position. Second, given that the extensions are defined at the level of type-shift rankings and DKP, they are intended to be language independent and the predictions for Hindi BSs should extend to BSs in any other articleless language. Finally, under the assumption that articleless languages without a grammaticalized singular/plural distinction in the nominal domain do not impose restrictions on the application of DKP (Dayal, 2004, 430), Dayal predicts them to differ from languages like Hindi and always allow for indefinite readings of their BNs. In what follows, we present Dayal’s take on these predictions for Hindi, Russian and Mandarin and discuss how they have been received in the literature.

For Hindi, Dayal admits that there are cases in which BSs seem to get an indefinite reading (see Dayal, 2011) :

- (4) anu **kitaab** paRhegi
 Anu book read-FUT
 ‘Anu will read a book.’

To account for cases like (4), Dayal argues that *kitaab* does not appear in regular argument position but rather in a pseudo-incorporated position. Crucially, pseudo-incorporated nouns can be argued not to type-shift, their apparent indefiniteness stemming from the construction they appear in. As such, examples like (4) do not need to pose a threat for Dayal’s prediction that Hindi BSs in regular argument position only take on definite readings.

Dayal takes Russian to be a good example of another articleless language with a grammaticalized singular/plural distinction in the nominal domain and argues that Russian BSs align with their Hindi counterparts. Example (5) replicates the BS/BP asymmetry we saw in (1) (see Dayal, 2004):

- (5) a. #**Sobaka** byla vezde
 dog was everywhere
 ‘The same dog was everywhere.’
 b. **Sobaki** byli vezde
 dogs were everywhere
 ‘There were dogs everywhere.’

Whereas (5b) straightforwardly allows for the reading according to which there were dogs everywhere, the singular *sobaka* only leads to the same pragmatically odd reading as (1a), according to which the same dog was everywhere.

For articleless languages that allow for BNs but do not have a grammaticalized singular/plural distinction, Dayal discusses Mandarin and points out that Mandarin BNs are on a par with Hindi BPs rather than with Hindi BSs in allowing for narrow scope readings in contexts like (6) (see Dayal, 2004):

- (6) **gǒu** zài měi gè rén de hòu yuàn li jiào.
 dog at everyone-DE backyard-inside bark
 'Dogs were barking in everyone's backyard.'

Example (6) is compatible with a reading in which different dogs are barking in different people's backyards. This reading is similar to the one we get for Hindi BPs in (1b), in line with Dayal's prediction.

In the formal semantics literature, Dayal's account has been the predominant one for Hindi BNs and the literature on Mandarin has not called into question the main prediction Dayal makes. For Russian, the story is different and multiple authors have argued that Russian BSs do not show any signs of inherent definiteness (e.g., Bronnikov, 2006; Šimík and Demian, 2020; Seres and Borik, 2018). Example (7) illustrates this (see Seres and Borik, 2018):

- (7) V každom dome igral **reběnok**.
 in every house played child.NOM

Example (7) straightforwardly allows for a reading according to which different children were playing in different houses, showing that the BS *reběnok* can take narrow scope under the universal *každy dom*. We admit that the structure of (7) is possibly different from the one in (5a) but this should not affect Dayal's prediction, and we conclude that (7) constitutes a clear counterexample.

3.2.3 Towards a cross-linguistic re-assessment of Dayal's account

Although the Russian facts have an immediate impact on the validity of Dayal's analysis, we are not aware of any attempt at re-evaluating Dayal's account for other languages than Russian. We assume that this is because the literature – up till recently – lacked the right tools to compare the distributions of BNs across languages and properly assess the empirical scope of counterexamples like (7). In this paper, we propose a translation corpus study and assess the predictions Dayal makes by analyzing translations of the same text to Hindi, Russian and Mandarin, allowing for a broad parallel evaluation of Dayal's predictions for these three languages.

3.3 Methodology

Translation corpus research has recently been argued to constitute a valuable addition to the toolbox of semanticists who study cross-linguistic variation. The phenomena that have been studied include – among others – tense and aspect (Fuchs and González, 2022; van der Klis et al., 2022; Mulder et al., 2022; de Swart et al., 2022a; Tellings and Fuchs, 2021), negation (de Swart et al., 2022b) and reference (Bremmers et al., 2022). As for languages, the main focus has been on Romance and Germanic, but we also find extensions to a broader set of European languages (Gehrke, 2022; de Swart et al., 2022b) as well as to Mandarin (Bogaards, 2022; Bremmers et al., 2022; Mo, 2022). Parallel to theory-oriented research, recent work is also covering the methodological side of translation corpus research from a semantics perspective (Le Bruyn et al., 2022; Le Bruyn and de Swart, 2023, 2024).

The main advantage of translation corpora is that they present the same semantic content in a maximally similar way in different languages. For research into reference, this means that we can neatly trace the ways different languages deal with reference in the same (or maximally similar) contexts. By choosing a source language that makes a formal distinction between definiteness and indefiniteness, we can furthermore use this distinction as an independently motivated criterion to distinguish between definite and indefinite reference in languages that have been argued not to mark this distinction.

The source corpus we selected is the first chapter of *Harry Potter and the Philosopher’s Stone*, a fairly recent novel that has been translated to Russian, Hindi, and Mandarin but also to an impressive array of other typologically diverse languages, allowing for the easy scaling up of the approach we pursue. We extracted all referential expressions from the English source text ($n = 1210$) but for the current research, we focus on $a(n)+N_{sg}$ ($n = 90$), $the+N_{sg}$ ($n = 140$) and N_{pl} ($n = 52$) and look into how they are rendered in the Russian, Hindi and Mandarin translations of the novel. The choice of these referential expressions is inspired by the dimensions that play a role in Dayal’s analysis: number and (in)definiteness.

For $a(n)+N_{sg}$, Dayal predicts BS translations to be rare in Hindi and Russian and for BN translations to be perfectly fine in Mandarin. For Hindi, BSs should only be allowed to occur in pseudo-incorporation constructions (as in (4)) whereas, in other contexts, the Hindi translator is predicted to rely on overt determiners, the \exists -shift and DKP both being cut off as viable paths to indefiniteness for BSs. In line with the examples Dayal presents herself, the default way of rendering a singular indefinite in

Hindi is to rely on *ek* ('one'):

- (8) *bahut saal pahle, yehaaN *(ek) aurat rahtii thii.*
 many years ago, here *(one) woman lived
 'Once upon a time, a woman used to live here.'

For Russian, we expect to find the same empirical picture as for Hindi, BSs being the minority option and determiners like *odin* ('one') being the default option for rendering singular indefinites. Given that Dayal does not cut off the DKP path to indefiniteness for Mandarin BNs, she predicts the latter to be viable translations for singular indefinites.

For *the+N_{sg}* and for *N_{pl}*, Dayal predicts BNs/BSs/BPs to be the default options in all three languages. Given that she ranks the ι -shift at the same level as the \cap -shift, each of the languages should straightforwardly allow its BSs/BNs to appear in singular definite contexts. Furthermore, given that Dayal takes DKP not only to be a viable path to indefiniteness for Mandarin BNs but also for Hindi and Russian BPs, we expect to find BPs/BNs as the default translations of *N_{pl}* in all three languages.

In Section 3.4, we organize the presentation of the results around the three types of contexts we have sketched above: singular indefinites, singular definites and plural indefinites. Because of this division of contexts, we can abstract away from number marking in Russian and Hindi, allowing us to resort to BNs as a general label and directly compare our Russian, Hindi, and Mandarin data. For each of the contexts, we compare the three languages and present descriptive and—where applicable—inferential statistics. The inferential statistic we will rely on is Fisher's Exact Test, an alternative to the classic chi-square test that provides more reliable results for smaller datasets in which some expressions are far less frequent than others.

One final remark is in order before turning to the results. Even though translations render the same meaning as their original texts, it does happen that translators opt for different structures in which the referents of the original are not translated one-on-one. A concrete example from our corpus is *having a tantrum* that is translated to Mandarin as *fā pīqì* (lit. "lose temper"): the overall meaning is the same but there is no direct reference to a tantrum in the translation. We separately report on these cases but do not take them into account in our analyses.

3.4 Results

3.4.1 Singular indefinite contexts

For singular indefinite contexts ($n = 90$), we found 23 cases of different constructions in Mandarin, 9 in Russian and 6 in Hindi. We report on three types of translations: (i) BNs, (ii) numeral-one+N, (iii) rest. For Hindi and Russian, BNs are restricted to BSs and for Mandarin, the numeral option includes a classifier. Figure 3.1 summarizes the data.

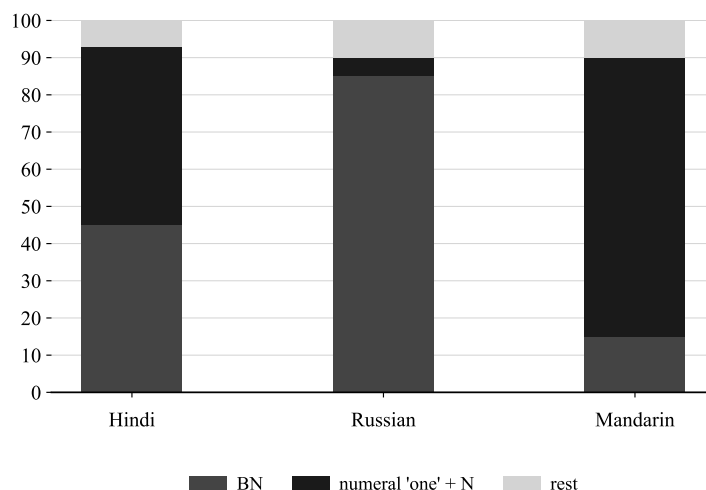


Figure 3.1: Relative frequencies of BN, numeral-one+N, and other translations of English indefinite singulars ($a(n)+N_{sg}$) in Hindi, Russian, and Mandarin.

Figure 3.1 shows that there are big differences in how each language renders singular indefinites. Whereas Russian barely relies on the numeral, the latter is slightly more frequent than the BS in Hindi and is clearly the majority option in Mandarin. The differences in distribution of BNs and the numeral are also statistically significant ($\alpha = .05$), Fisher’s Exact Tests leading to p -values smaller than 0.01 for the comparisons of the different language pairs. The rest category is varied in each of the languages but BNs and numeral-one+N clearly come out as the majority options for Hindi and Mandarin. In Russian, none of the rest options (proper names, pronouns, indefinite determiners, etc.) appear in more than two contexts.

3.4.2 Singular definite contexts

For singular definite contexts ($n = 140$), we found 18 cases of different constructions in Russian, 12 in Mandarin and 5 in Hindi. Across the three languages, there was one construction that—despite remaining a distant second overall—stood out: the demonstrative. Even though Dayal makes no explicit predictions about the competition between BNs and demonstratives, our data do suggest that there is an interaction between the two and we consequently report on (i) BNs, (ii) demonstrative + N, (iii) rest. As for singular indefinites, BNs are restricted to BSs for Hindi and Russian. For Mandarin, the “demonstrative + N” option typically contains a classifier. We summarize the data in Figure 3.2.

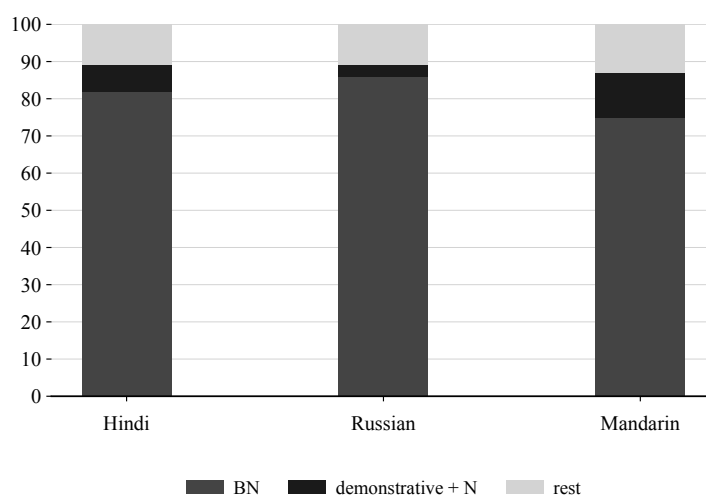


Figure 3.2: Relative frequencies of BN, demonstrative + N and rest translations of English definite singulars (*the* + N_{sg}) in Hindi, Russian and Mandarin.

Figure 3.2 shows that BNs are the majority option in all three languages. At the same time, we see that demonstratives are gaining ground, in particular in Mandarin. Pairwise comparisons between the languages show that the differences in distribution of BNs and demonstratives are significant for Russian-Mandarin ($p < 0.01$, Fisher’s Exact Test) but not for Russian-Hindi ($p = 0.15$) nor for Hindi-Mandarin ($p = 0.14$).

3.4.3 Plural indefinite contexts

For plural indefinite contexts ($n = 52$), we found 5 cases of different constructions for Russian, 4 for Mandarin and 4 for Hindi. No constructions involving plural determiners appeared in more than two contexts in any of the languages, leaving us with no clear competitors to compare BNs to. In the absence of obvious competitors, we refrain from presenting graphs with relative frequencies and running inferential statistics. Our data show that BNs/BPs come out as the main category for translating plural indefinites in all of the languages ($n = 31$ in Hindi, $n = 32$ in Russian, $n = 39$ in Mandarin).

3.5 Discussion

In Section 3.2 and Section 3.3, we worked out the predictions Dayal makes for the translation of singular indefinites, singular definites and plural indefinites to Hindi, Russian and Mandarin. For singular definites and plural indefinites, we argued that Dayal predicts BNs/BSs/BPs to be the default options. For singular indefinites, however, Mandarin would have BNs as the default option whereas Hindi and Russian should both show a clear preference for nouns preceded by indefinite determiners like the numerals *ek* and *odin* (“one”).

The picture that emerges from our results in Section 3.4 is different from the one predicted by Dayal. In this section, we go through the different contexts, discuss in how far our data are in line or at least compatible with Dayal’s predictions and explore extensions and modifications where relevant. Throughout, we will argue that Dayal’s analysis has to be extended and ultimately modified. The alternative analysis we move towards is one in which so-called “articleless” languages do have articles that compete with BNs in varying ways. We start with singular definite and plural indefinite contexts, keeping singular indefinite contexts—the most problematic case—for last.

3.5.1 Singular definite contexts

Our singular definite data are overall in line with Dayal’s predictions in the sense that BSs/BNs clearly constitute the majority option in all three languages. The one surprise in our data is the special role of demonstratives that leads to a statistically traceable difference between Russian and Mandarin. Given that Hindi demonstratives do not lead to significant differences with Russian or Mandarin, we focus here on the Mandarin case.

The role of demonstratives in the referential system of Mandarin is not predicted by Dayal in her 2004 paper but has received attention in the more recent literature. Jenks (2018) argues that Mandarin demonstratives function as grammaticalized markers of familiarity and block BNs from marking this subtype of definiteness. In what follows, we argue that the division of labor between BNs and demonstratives is different from the one proposed by Jenks but does not jeopardize the core of Dayal's analysis.

Our data show that demonstratives are used in familiarity contexts (10), but at the same time, we find that they are not obligatory in these contexts (9), contra Jenks (2018). Both (9) and (10) are part of a bigger context in which a cat is introduced and referred back to, (9) occurring before (10).

- (9) Mr Dursley blinked and stared at the cat. It stared back.

Désǐlǐ xiānsheng zhǎ le zhǎ yǎn, dīng zhe **māo** kàn
Dursley Mr blink LE blink eye stare ASP cat look

- (10) [...] he watched the cat in his mirror.

tā cóng hòushìjìng lǐ kànkàn **nà zhī māo**.
he from rear-view-mirror inside look that CL cat

Both *māo* in (9) and *nà zhī māo* in (10) refer back to the same cat that was introduced earlier. They thus count as familiar definites and show that Mandarin resorts both to BNs and to demonstratives in familiarity contexts. The exact division of labor between the two is an empirical puzzle that has been tackled in several recent papers (Bremmers et al., 2022; Dayal and Jiang, 2022; Simpson and Wu, 2022). The data in (9) and (10) are in line with Bremmers et al. (2022)'s proposal that Mandarin is sensitive to situation-level familiarity, allowing for familiar readings of BNs if they are introduced in the same situation as their antecedent and requiring the use of demonstratives to refer back to referents introduced in different situations. We refer the reader to Bremmers et al. (2022) for further details but the intuition for (9) is that it is part of the same scene in which the cat is introduced through the eyes of Mr Dursley whereas (10) is part of another scene in which Mr Dursley drives off to work and looks back at the cat through his rear-view mirror. In line with Bremmers et al. (2022), we find that the BN can felicitously refer back to the cat within the same scene it was introduced in but that the translator resorts to the demonstrative when referring back to the cat in a separate scene.

Clearly, further research is needed to unpack the Mandarin data further and compare the different proposals on the division of labor between BNs and demonstratives.

Crucially, though, our data suggest that such a division of labor exists and argue in favor of analyzing Mandarin demonstratives as article-like expressions that compete with BNs. Dayal does not foresee a role for definite articles in Mandarin but adding them does not impact on the core of her analysis: in neo-Carlsonian analyses, articles constitute an additional layer that is independent of the rankings of type-shifts and of number constraints on DKP. If we extend Dayal’s analysis with the assumption that Mandarin demonstratives function as articles, the prediction that follows is that BNs freely take on definite readings except for the subtype that demonstratives specialize in. This prediction is in line with our data, and we conclude that singular definite contexts do not pose a threat to Dayal’s analysis.

3.5.2 Plural indefinite contexts

With English N_{pl} being predominantly translated as BNs/BPs in Hindi, Russian, and Mandarin, our plural indefinite data align with Dayal’s prediction that BNs/BPs should straightforwardly give rise to indefinite readings in all three languages. As we indicated in Section 3.4.3, it is important to unpack the data further and comment on three tendencies in the source and target languages. These tendencies highlight the need to always run both quantitative and qualitative analyses, but we will argue that none jeopardize Dayal’s analysis.

Kind readings

When analyzing the N_{pl} data, we noticed that not all convey indefinite readings; four turn out to convey kind readings. Example (11) is a representative example:

- (11) [English:] Although **owls** normally hunt at night and are hardly ever seen in daylight, there have been hundreds of sightings of these birds flying in every direction since sunrise.

In (11), *owls* does not refer to an indefinite plurality but to owls in general and counts as kind-referring. These kind-referring instances show that our form-driven approach must be complemented with fine-grained semantic analysis. The question is whether translations of these kind-referring N_{pl} align with Dayal’s predictions.

Sections 3.2 and 3.3 did not separately develop Dayal’s predictions for kind-referring N_{pl} , but they are straightforward: following the neo-Carlsonian premise that all BNs refer to kinds at some point in their derivation, Dayal predicts BNs/BSs/BPs to be viable translations. We argue that this is borne out.

In Mandarin, the standard translation is with a BN, except for one pronoun case in (12):

- (12) ‘Normally, **they** [owls] prey at night and rarely show themselves during the day, but today, at sunrise, owls were flying all over the place.’

tōngcháng qíngkuàng xià, **tāmen** dōu shì zài yèjiān bǔshí, báitiān
 normally situation under, they all be at night-time prey, day-time
 hěn shǎo lòumiàn, kěshì jīntiān, rì chū shí māotóuyīng jiù
 very rarely show-face, but today, sun rise moment owl then
 sìchù fēnfēi.
 all-around fly-fly.

Despite constructional differences, Mandarin translations point in a single direction: kind-referring English N_{pl} are translated with BNs. Hindi uses BPs; Russian, in three of four cases, uses BPs, with one using a plural noun plus *takiye* (‘such’).

We conclude that our kind-referring data align with Dayal’s predictions, consistent with our plural indefinite data.

Mandarin *-men* and functional readings

A second tendency involves the plural marker *-men* in Mandarin translations ($n = 3$). Though low-frequency, this is surprising since *-men* often relates to definite-like interpretations (Jiang, 2017). Closer scrutiny suggests that with *-men*, the translator picks up a third reading of English N_{pl} : the functional reading. Example (13) illustrates:

- (13) ‘**Experts** are unable to explain why the owls have suddenly changed their sleeping pattern.’

zhuānjiā-men yě wúfǎ jiěshì māotóuyīng wèishéme gǎibiàn-le tāmen
 expert-MEN also unable explain owl why change-ASP they
 de shuìmián xíguàn
 DE sleeping habit

Here, *zhuānjiāmen* translates English *experts* with a functional reading, referring to relevant bird experts. As with kind readings, functional readings highlight the need for combining form-driven and semantic analysis. Unlike kind readings, we cannot check if Dayal’s predictions cover functional readings, as she has not analyzed these explicitly. These data deserve further unpacking but do not directly challenge Dayal’s core predictions.

From English N_{pl} to Hindi nouns unmarked for number

The final tendency is that twelve English plural indefinites are translated with nouns unmarked for number in Hindi, a striking contrast to four such cases in Russian. This may stem from Hindi’s allowance of pseudo-incorporation, as in (14a), and from nouns like *dril* (‘drill’) that cannot be pluralized, as in (14b).

- (14) a. ‘He’d forgotten all about the people in cloaks [...]’
 Choga pahane **logo**-ke baare.mein ve pooree tarah bhool
 cloak wear.PERF people-GEN about they complete manner forget
 chuk-e the [...]
 PERF PST [...]
- b. ‘He found it a lot harder to concentrate on drills that afternoon [...]’
 us dopahar **dril** par dhyaan lagaaye.rakhane mein unhen bahut
 that afternoon drill on attention keep-continue in to.them much
 kathinaee huee [...]
 difficulty happened [...]

These cases suggest that Hindi’s use of unmarked nouns relates to pseudo-incorporation and number morphology issues. They have little bearing on the referential potential of Hindi and Russian BSs/BPs and do not challenge Dayal’s predictions.

3.5.3 Singular indefinite contexts

Up till now, we have argued that our data in singular definite and plural indefinite contexts align with Dayal’s predictions, with only the hypothesis that Mandarin demonstratives act as specific definite articles. Singular indefinite contexts, however, lead to both extensions and modifications.

Dayal predicts that singular indefinites translate as BNs in Mandarin, while Hindi and Russian prefer nouns preceded by indefinite determiners like *ek* or *odin*. Our Hindi data align with this; Mandarin and Russian do not. We illustrate using the translations of *a map* (15) and *a new word* (16).

- (15) “It was on the corner of the street that he noticed the first sign of something peculiar—a cat reading **a map**.”
- a. **Hindi:**
 Sadak-ke mod par Dursley ko pehli ajib chiz dikh-i —ek
 Street-GEN corner on Dursley to first strange thing.F see-PST.F —a
 billi, jo **naksha** padh rahi thi
 cat.F who map read PROG be.PST

b. **Russian:**

Tol'ko na uglu ulicy mister Dursley nakonec zametil, što
 only on corner street-GEN mister Dursley finally noticed, that
 proisxodit čto-to strannoe, –a zametil on košku, vnimatel'no
 happens something strange, and noticed he cat-ACC, attentively
 izučavšuju ležaščuju pered nej **kartu**
 examining lying in.front.of her map-ACC

c. **Mandarin:**

zài jiē.jiǎo shàng, tā kàn-dào-le dì-yī-gè yìcháng-de xìnghào
 at street.corner on, he see-RVC-ASP ORD-one-CL peculiar-DE sign
 — yì-zhī-māo zài kàn **dìtú**
 — one-CL-cat PROG read map

- (16) “She told him over dinner all about Mrs Next Door’s problems with her daughter and how Dudley had learnt **a new word** (‘Shan’t!’).”

a. **Hindi:**

Unho-ne dinner par apne pati ko bata-ya ki padosan ki apni
 she-ERG dinner on her husband to told-PFV that neighbor of own
 beti ke.sath kya samasyaye chal rahi hai aur Dudley-ne
 daughter with what problems go PROG be.PRES and Dudley-ERG
ek naya vakya sikh-a hai “nahi karu-n-ga”
 a new sentence learn-PFV be.PRES 'no do-FUT-M'

b. **Russian:**

Za obedom ona oxotno spletničala, rasskazav misteru Dursley
 at lunch she gladly gossiped, having.told mister-DAT Dursley
 o tom, što u ix sosedki ser'ěžnye problemy s dočer'ju, i
 about that, that at their neighbour serious problems with daughter, and
 naposledok soobščiv, što Dudley vjučil **novoe slovo** “xačču!”
 finally having.informed, that Dudley learnt new word “I.wanna!”

c. **Mandarin:**

wǎnfàn zhuō shàng, Désiǐ taitai xiàng tā jiǎngshùle línjū jiā
 dinner table on, Dursley Mrs to he tell-ASP neighbour family
 de mǔ-nǚ máodùn, hái shuō Dáli yòu xuéhuì **yíge**
 DE mom-daughter conflict, also say Dudley again learn-RVC one-CL
xīncí (“jué bù”)
 new-word (never)

The translations of *a map* and *a new word* neatly illustrate the two major patterns that emerge from our data: one in which the Mandarin and Hindi translators both opt

for a BN/BS (*a map*) and one in which they both opt for a construction with a numeral, Hindi *ek* and Mandarin *yi* “one” (*a new word*), the Russian translator choosing a BS in both cases. We comment on the third pattern—one in which Hindi and Russian opt for a BS but Mandarin resorts to a construction with *yi*—in due course. The attentive reader will have noticed that the English original in (15) also contains a second indefinite singular—a cat. We leave it aside as the structures of its translations vary slightly.

Hindi

For Hindi, our singular indefinite data overall seem in line with Dayal’s predictions, especially if we compare the frequency of BNs as translations of singular definites (over 80%) to the frequency of BNs as translations of singular indefinites (below 40%). Examples (15) and (16) nicely illustrate the alternation between BSs and nouns preceded by *ek* that we find in our singular indefinite data. To be fully in line with Dayal’s predictions, the argument should be that *naya vakya* occurs in regular argument position and therefore requires *ek*, but that *naksha* is pseudo-incorporated and can therefore occur without the numeral. Parallel to *kitaab* in (4), we assume that a pseudo-incorporation analysis for *naksha* is not implausible. We do note that the literature on pseudo-incorporation in Hindi does not give us any direct way to argue for it. In future work, we count on developing Le Bruyn et al. (2016)’s analysis of pseudo-incorporation and on exploiting the constraints it predicts on verb-noun combinations. According to Le Bruyn et al., pseudo-incorporation – in languages that allow for it – is possible in case the verb taps into the explicit or implicit relational semantics of the noun it combines with. This analysis gives us a handle on cases like *read book*, *read map* and *learn new word*. Indeed, whereas *book* and *map* both come with a telic role in their qualia structure (Pustejovsky, 1995) and can thus be argued to come with an implicit use relation that can be picked up on by *read*, a noun like *word* arguably does not come with any explicit or implicit relational semantics that *learn* can pick up on. The prediction Le Bruyn et al. (2016)’s analysis makes then is that *read book* and *read map* allow for pseudo-incorporation and that *learn (new) word* does not. These predictions are in line with the data in (4), (15) and (16), *read book* and *read map* leading to BS translations and *learn new word* leading to a translation with *ek*. We submit that a full analysis of the Hindi data requires further theoretical and empirical work but conclude that on the basis of the Hindi singular indefinite data alone, we have no reason to believe that they are incompatible with Dayal’s analysis. We get back to this

conclusion when we discuss the Russian indefinite singular data (Section 3.5.3).

Mandarin

For Mandarin, Dayal's prediction is that BNs should straightforwardly give rise to indefinite interpretations. With BNs occurring as translations of singular indefinites in fewer than twenty percent of the cases, this is arguably not the empirical picture we find. A counterargument one can entertain is that the low frequency of BNs in Mandarin does not say anything about the grammaticality of indefinite interpretations of BNs, but this argument makes little sense from Dayal's perspective if the low frequency of BSs in Hindi is to be indicative of their ungrammaticality in regular argument position. We argue that our singular indefinite data are not in line with the predictions Dayal makes and that the analysis she proposes for Mandarin has to be adapted.

There are at least two options available to make sure that Mandarin BNs are not predicted to freely occur in singular indefinite contexts. One is to reconsider Dayal's assumption that DKP is freely available for BNs in Mandarin. The disadvantage of this strategy is that it would make the prediction that BNs also have a hard time getting an indefinite interpretation in plural contexts, contrary to fact (see Section 3.4.3). The other option is to assume that Mandarin is not only developing a definite article (see Section 3.5.1) but also an indefinite one, specifically in the singular domain. Unlike the DKP strategy, the article strategy correctly targets the singular domain alone. It does raise the question what it means to be a developing indefinite article and how we can best formalize the division of labor between BNs and this indefinite article in synchrony. In this respect, a relevant tendency in our data is that there is only one case of a BN in Mandarin that is translated with the numeral in Hindi whereas there are fifteen cases of Hindi BSs that are translated with the numeral in Mandarin. What this tendency suggests is that Mandarin BNs are more restricted than Hindi BSs but that they do share a common set of contexts in which they appear. One route that deserves to be explored then is that Mandarin *dítú* in (15) is pseudo-incorporated in the same way as Hindi *naksha* and that the division of labor between the Mandarin developing indefinite article and Mandarin BNs is to be formalized as a competition between the indefinite article and BNs in pseudo-incorporation constructions.

We conclude that our data point to the need to adapt Dayal's analysis for Mandarin and that extending it with the hypothesis that Mandarin is developing an indefinite article holds promise. We furthermore conclude that our data are suggestive of

a scalar relation between contexts allowing for BSs in Hindi and BNs in Mandarin, arguing in favor of analyzing the division of labor between the Mandarin indefinite article and Mandarin BNs as a competition between the article and BNs in pseudo-incorporation constructions. Here too, we hope to develop Le Bruyn et al. (2016)’s analysis of pseudo-incorporation in our future work, as their analysis is explicitly set up in terms of a competition with singular indefinite articles.

Russian

For Russian, Dayal’s predictions are similar to the ones for Hindi, BSs being clearly dispreferred and the translator opting for nouns preceded by an indefinite determiner like *odin* in the great majority of the cases. For Hindi, our data were overall in line with these predictions, but our Russian data show a completely different picture: unlike in Hindi, BSs are by far the predominant option to render singular indefinites in Russian. Examples (15) and (16) are representative examples: where Hindi alternates BSs and nouns preceded by *ek*, Russian uniformly opts for BSs. We conclude that our Russian indefinite singular data are not in line with Dayal’s predictions.

To accommodate our Mandarin singular indefinite data, it sufficed to extend Dayal’s analysis with the hypothesis that Mandarin is developing an indefinite article. To accommodate our Russian data, we do not see how a simple extension could make do. The problem is that Dayal has meticulously closed off all routes to indefinite readings for regular BS arguments and those are exactly the ones we need. At the same time, it seems that we cannot re-open these routes without making the wrong predictions for Hindi. We are thus faced with a stalemate: either we get the Russian data right and the Hindi data wrong or vice versa.

To break the stalemate, we think it is instructive to zoom out and go back to the type of examples that originally motivated Dayal’s analysis, *viz.*, those in which Hindi BSs turn out to resist narrow scope indefinite interpretations (see (1a)). Crucially, we find the same resistance to narrow scope with unambiguously indefinite expressions like English $a + N_{sg}$:

(17) **A dog** was everywhere.

As noted by Carlson (1977), (17) only has the same bizarre reading as *cuuha* in (1a), *viz.*, one in which the same animal is said to be everywhere. What this suggests is that the odd reading of (1a) is unlikely to be due to a definite interpretation of *cuuha* and that closing off the existential and the DKP route to indefinite readings for BSs is

not offering a solution to the real puzzle (1a) raises, *viz.*, one that is not concerned with the absence of indefinite interpretations of BSs but with missing narrow scope readings of indefinite singulars. Even though we will not try to attempt to solve the real puzzle, we submit that there are information-structural considerations at play, explaining why minor variations on the same sentence lead to different intuitions ((5a) vs. (7)).

Under the assumption that Dayal's closing off of the existential and the DKP route is the wrong way to derive the odd reading of (1a), we argue that at least one of the routes should be reopened, either deciding that DKP can apply to singular kinds or returning to Chierchia's original type-shift ranking. The upshot of this move is that—independently of our pick—our Russian data follow straightforwardly.

With the Russian data accounted for, the last step to be taken is to offer an alternative account for the fact that Hindi BSs in our corpus give rise to indefinite readings far less easily than in Russian. We hypothesize that Hindi, like Mandarin, is developing an indefinite singular article that competes with BNs in pseudo-incorporation constructions. Unlike the blocking of DKP for singular kinds and the re-ranking of type-shifts, the hypothesis of a developing indefinite article can be done at a language-specific level and—as such—allows us to account for the fact that BSs have a hard time getting indefinite readings in Hindi while at the same time making sure that they freely get these readings in Russian. We conclude that our article strategy allows us to break the stalemate we faced. The fact that articles can show different degrees of grammaticalization furthermore comes with the additional perk of creating the flexibility we need to account for the difference in distribution of Hindi and Mandarin numerals, another theoretical and empirical challenge we, however, have to leave for future work.

3.5.4 Recap

Throughout this section, we have argued that our data are compatible with many of Dayal's predictions but that her analysis does go wrong on some crucial points, in particular for definite singular contexts in Mandarin (Section 3.5.1) and for indefinite singular contexts in Mandarin and Russian (Section 3.5.3). To accommodate the Mandarin data, we argued that it suffices to extend Dayal's analysis, and we hypothesized that Mandarin is developing an indefinite and a definite article. Accommodating the Russian data turned out to require a real modification of Dayal's analysis, re-opening at least one of the two paths to indefinite readings for BSs that Dayal meticulously closed off. With the Mandarin and Russian data accounted for, we were left with the Hindi data that originally motivated the closing off of the indefiniteness paths for BSs.

Given that there was no way to accommodate the Russian data otherwise, we proposed an alternative analysis for Hindi, hypothesizing that—parallel to Mandarin—it is developing a singular indefinite article. Further theoretical and empirical work is needed but we do believe we have laid the necessary groundwork to build on in our future work.

The picture that has emerged throughout this section is that some so-called “articleless” languages are less articleless than the literature has assumed up till now. We found that Hindi, Russian and Mandarin behave truly differently from each other and that capturing these differences requires us to abandon language-independent strategies to account for language-specific tendencies. By resorting to the hypothesis that Mandarin and Hindi are developing articles, we opted for a language-specific strategy that holds the promise of capturing the tendencies we found in our Mandarin and Hindi data without generating predictions that pose a problem for Russian, in which BSs really do turn out to freely allow for both definite and indefinite readings.

3.6 Conclusion

In this paper, we adopted a translation corpus approach based on the first chapter of *Harry Potter and the Philosopher’s Stone* to come to a broad parallel evaluation of Dayal’s seminal work on reference in Hindi and the predictions it makes for Russian and Mandarin. Dayal’s core intuition is that Hindi BSs are different from Hindi BPs in that they do not allow for indefinite readings. In Section 3.2, we worked out how Dayal accounts for this intuition by closing off the two paths to indefinite readings that Chierchia’s original version of the neo-Carlsonian framework left for BSs in articleless languages. In Sections 3.3 to 3.5, we worked out how Dayal’s predictions can be operationalized for translation corpus research and argued that our Hindi data are overall compatible with them but that the same does not hold for our Mandarin and Russian data, leading us to explore a number of extensions and modifications of Dayal’s analysis. For Mandarin, our data led us to hypothesize a role for the numeral as an indefinite article and for demonstratives as definite articles. For Hindi and Russian, we argued that the only way to account for the two languages was to re-open at least one of the two paths to indefinite readings Dayal closed off and to hypothesize that Hindi—unlike Russian—is developing an indefinite article. The overall conclusion we arrive at is that so-called “articleless” languages are not created equal.

Along the way, we pointed out that there remains quite some theoretical and empirical work, in particular to properly pin down what it means for a language to be

developing a definite and an indefinite article (see Liu et al., 2022 and Bremmers et al., 2022 for some first steps). Relevant follow-up empirical work also includes replication and triangulation of our results as well as extensions to a broader set of languages (see Borik et al., 2025). At a more general theoretical level, it is important to assess the impact of our data on the neo-Carlsonian framework, paying special attention to the desirability of reversing Dayal's extensions and how these would fare with later updates of the framework (see Liu et al., 2023a for discussion).

CHAPTER 4

The Theory of Argument Formation: between Kinds and Properties¹

4.1 Introduction

Chierchia's (1998) cross-linguistic extension to Carlson's (1977) kinds analysis of English bare plurals (BPs) (henceforth the *Kinds Approach* (KA)) is the most influential theory of argument formation to date. Among the core facts that it was able to derive, we mention the generalized narrow scope behavior of bare nouns (BNs) and the existence of generalized classifier languages. After Chierchia (1998), the KA was further developed and a number of competing theories were proposed. None of them, however, have achieved anything close to the same popularity.

We single out Krifka (2003) as the KA's conceptually closest competitor. Both the KA and Krifka's approach are cast in a type-shifting framework along the lines set out

¹*Chapter 4 was originally published with the title The theory of argument formation: between kinds and properties in the proceedings of SALT 33 and is joint work with Shravani Patil, Hagay Schurr, Daria Seres, Olga Borik and Bert Le Bruyn. The original paper is positioned as an exploration of the options the Kinds and the Properties Approach have to account for data from a typologically broad range of languages. This chapter slightly adapts the exploratory theoretical perspective from the original paper and restricts it to Mandarin, Hindi and Russian to fit its role in the thesis argumentation. See in Section 1.5 for more details on the argumentation overview of this chapter.*

in Partee (1987). However, while Krifka's approach takes nouns to uniformly start life as predicates (hence we qualify it as a *Properties Approach* (PA)), the KA does not extend this to all languages, as we discuss in Section 4.2.1. The other crucial difference between the two lies in the fact that predicates-to-arguments shifts are not ranked in the PA. The PA consequently does not attribute any special status to predicates-to-kinds shifts, differently from the KA.

Two developments related to the core facts mentioned above invite an open-minded reassessment of the explanatory potential of the KA and the PA.

On narrow scope

Even though the narrow scope behavior of BNs is often considered to require an approach with a central role for kinds, Krifka (2003) was one of the first to argue that all one needs is a locality requirement on type-shifting, an assumption that is also central to the KA. The exact implementation of the locality requirement is slightly different in the two approaches, though. Comparing the narrow scope accounts of both approaches, Le Bruyn and de Swart (2022) find that they differ in their predictions about the scope of scrambled BNs. Whereas the KA predicts no difference between scrambled and unscrambled BNs, the PA predicts the former to be able to take wide scope. Le Bruyn and de Swart (2022) argue that scrambled bare plurals in Dutch take wide scope over negation and consider this to be an empirical argument in favor of the PA.

On classifier languages

One of the most appealing achievements of Chierchia's original version of the KA was that the existence of generalized classifier languages like Mandarin followed directly from the notion that some languages can have their nouns uniformly start life as kinds. Krifka was able to mimic the need for classifiers in Mandarin but needed to do so in the lexical entries of individual nouns, a *prima facie* less attractive move. However, after developmental psychologists and linguists had criticized the original version of the KA for ignoring the distinction in Mandarin between mass and count nouns, Chierchia (2010) was led to refine his theory and assume that nouns are lexically marked to start life as mass or count kinds (see also discussion in Jiang (2020)). The crucial point here is that the KA does not escape the need to adopt the same type of lexicalist approach as the PA, arguably leveling the playing field between the two.

Since these developments suggest that two major achievements of the KA have not fully withstood the test of time, we find that the PA comes out as a relevant competitor,

which calls for an open-minded reassessment of the predictions of both approaches.

Methodologically, we want to cast the net wide, and we do so in two respects. First, we consider a sample of six languages. In addition to four that represent the same languages or language families that appeared in Chierchia (1998), *viz.* Spanish (Romance), German (Germanic), Russian (Slavic), and Mandarin (Sino-Tibetan), we added Hindi for its pivotal role in the KA (for an exploration of bare nominals in Hindi, an “article-less” language, and its possible implications for our understanding of kind reference, see Dayal, 2004)), and Hebrew as a distinct in-between language type with its definite but no indefinite article (Doron, 2003), which sets it apart from both “article-less” and article languages in the rest of our sample. Second, we do not focus on preset examples but rather rely on what the analysis of a (small) translation corpus brings us, *viz.* the translations of the first chapter of J.K. Rowling’s *Harry Potter and the Philosopher’s Stone* (henceforth *HP*). Translation corpus research has recently gained traction as a valuable tool in the cross-linguistic semanticist’s toolbox next to questionnaires and experimental methodologies (see, e.g., Bremmers et al., 2022; Mulder et al., 2022; Gehrke, 2022; van der Klis et al., 2022).

The increasing number of papers that make use of translation corpus methodology in cross-linguistic semantics attests to the gradual but steady maturation of this subfield (Le Bruyn et al., 2022, 2024; Le Bruyn and de Swart, 2024). In addition to the maximal semantic comparability of parallel translations, we find the translation corpus approach on the basis of *HP* particularly attractive for two additional reasons: (i) the English original gives us a reasonable grip on the interpretation of BNs in the translations, even for languages that do not overtly mark (in)definiteness; (ii) the corpus can easily be extended, both in number of languages (*HP* was translated to over 80 languages) and in number of words (*HP* is a 7-volume series that makes for a corpus of approximately one million words, if used in its entirety, and we are only looking into the first chapter of the first volume).

The paper is organized as follows. In Sections 4.2 and 4.3, we work out the predictions the KA and PA make for the availability of BNs in argument position in singular/plural (in)definite contexts. In Section 4.4, we present the predictions of the KA and PA to be tested against the results of our parallel corpus study. We also add a brief note on pseudo-incorporation. Section 4.5 concludes the paper.

4.2 The Kinds Approach and its predictions

The KA draws on Carlson's 1977 influential work on BPs in English for the basic intuition that kind and indefinite readings of BNs are related, a common thread throughout the history of the KA. Next, we work out the hypotheses that make up the KA. Given the approach typically distinguishes between classifier and non-classifier languages, we discuss these in turn in Section 4.2.1 and Section 4.2.2. In Section 4.2.3, we work out the predictions the KA makes for the availability of BNs in argument position in (in)definite singular/plural contexts in the languages of our sample.

4.2.1 Non-classifier languages

For non-classifier languages, we first sketch the approach proposed in Chierchia (1998) and then highlight the refinements that have been implemented since.

Chierchia (1998)

Chierchia (1998) proposes that non-classifier languages come in two guises, related to a parameter he calls the *Nominal Mapping Parameter*:

- [-arg,+pred]: nouns uniformly start life as properties and need to rely on overt or covert functional material to shift to argumental types;
- [+arg,+pred]: nouns start life as properties or kinds and can shift between argumental and non-argumental types without the intervention of overt or covert functional material.

Building on Partee (1987), Chierchia assumes a type-shifting framework in which nouns can flexibly shift between types. The basic type-shifts he assumes are: the *iota* shift (ι , from type $\langle e, t \rangle$ to type e), the existential shift (\exists , from type $\langle e, t \rangle$ to type $\langle \langle e, t \rangle, t \rangle$), the *down* shift (\sqcap , from type $\langle e, t \rangle$ to kinds, type e_k), and the *up* shift (\sqcup , from kinds to their instantiations). Chierchia hypothesizes that shifts to argumental types are constrained by the type-shift ranking in (1) and the Blocking Principle in (2), the former prioritizing the *down* shift over the *iota* and the existential shifts, the latter proscribing shifts from applying covertly in languages that have lexicalized the shifts in their determiner systems. Definite articles are typically thought to lexicalize the *iota* shift and indefinite articles the existential shift.

- (1) Type-shift ranking (to be adapted): $\sqcap > \{\iota, \exists\}$

(2) The Blocking Principle:

For any type shifting operation τ and any X : $*\tau(X)$ if there is a determiner D such that for any set X in its domain, $D(X) = \tau(X)$

Outside standard type-shifting, Chierchia assumes kinds can give rise to derived indefinite readings through *Derived Kind Predication* (DKP), as defined in (3):

(3) Derived Kind Predication (DKP):

If P applies to objects and k denotes a kind, then $P(k) = \exists x[^{\cup}k(x) \wedge P(x)]$

DKP is Chierchia's formalization of the link between kind and indefinite readings of BNs and is the key to deriving the generalized narrow scope behavior of BNs (see Section 4.1). In essence, it takes a kind k and the predicate P that the k combines with, and returns the proposition that asserts the existence of individuals that are members of the kind and satisfy the descriptive content of the predicate. DKP is assumed to apply locally, making sure that the existential quantifier it generates always takes the narrowest possible scope.

After Chierchia (1998)

Where Chierchia (1998) distinguishes two types of non-classifier languages, we find a reduction to a single type in the later literature, viz. $[-arg, +pred]$ (Dayal, 2004; Chierchia, 2010; Jiang, 2020). The original motivation for the distinction between two types came from the contrast between languages like Italian and English: the former restricts the syntactic positions in which its BPs can appear whereas the latter does not (see also Longobardi, 1994). However, the current consensus attributes this fact to syntax, essentially rendering redundant any attempt to capture it with a semantic parameter. For example, Jiang (2020) derives the Italian/English opposition from the syntactic parameter $\pm \text{ARG}_{\text{unrestricted}}$.

A further evolution lies in the ranking of type-shifts. Chierchia originally proposed that the *down* shift should be ranked above both the *iota* and the existential shift. Starting from Dayal 2004, the consensus seems to be that the *iota* shift should be unranked with respect to the *down* shift and that both should be ranked above the existential shift (see also Jiang, 2020), resulting in the final ranking in (4):

(4) Type-shift ranking (final): $\{\cap, \iota\} > \exists$

For languages with definite articles, this change does not affect the predictions made by the KA, as the Blocking Principle (2) independently blocks the *iota* shift for

BNs. For languages like Hindi, with no definite article, this change entails that nouns may directly undergo both the *down* shift – leading to a kind reading – and the *iota* shift – leading to a definite reading.

Lastly, researchers within the KA have increasingly worked out how number interacts with kinds. Chierchia (1998) already hypothesized that the *down* shift requires plural nouns. Dayal (2004) emphasizes that the same holds for the *up* shift (\cup) in number-marking languages and restricts it to kinds built from *plural* nouns. The *up* shift is a crucial ingredient of DKP. With the restriction of the *up* shift to plural kinds, DKP can give rise to indefinite readings for kinds built from plural nouns but not for kinds built from singular nouns.

4.2.2 Classifier languages: the case of Mandarin

The KA assumes that classifier languages are $[+arg, -pred]$ languages and that their nouns start life as kinds. As we indicated in Section 4.1, the KA has been refined in that it now recognizes both count and mass kinds. This refinement has no direct impact on the predictions the KA makes about the availability of BNs in argument position, though.

For Mandarin, the availability of BNs in argument position in definite and indefinite contexts is worked out in most detail in Jiang (2020). Definite readings of BNs are derived through *Situation Restriction* (SR), as defined in (5).² It takes a kind and returns the maximal member instantiating it in a situation s . Indefinite readings are derived through *Derived Kind Predication* (DKP), as defined in (3).

(5) Situation Restriction (SR):

$$[N_{\langle e_k \rangle}]_s \rightarrow [N_{\langle e \rangle}] = \text{the maximal member instantiating } N_{\langle e_k \rangle} \text{ in a situation } s$$

We note that Mandarin is a classifier language that does not mark number. Number considerations are consequently not assumed to play a role in the availability of DKP in this language.

4.2.3 Predictions

With the KA's hypotheses in place, we can work out the predictions this approach makes for the availability of BNs in argument position in (in)definite singular/plural contexts for the languages under investigation. We will do so in two steps, first working out the predictions for non-classifier and classifier languages in general, and then

²The reader is referred to Jiang (2020) for further details.

fine-tuning them for the languages in our sample based on the articles the languages are assumed to have.

The predictions we will arrive at disregard the possible occurrence of BNs in pseudo-incorporation constructions (we return to this in Section 4.4). In this section and in Section 4.3, we focus on standard argumental BNs, as we assume that they can be distinguished from pseudo-incorporated BNs in a principled way.

General predictions

Assuming together with the KA's proponents that BNs in non-classifier languages start life as type $\langle e, t \rangle$ expressions and that they can be singular or plural, BNs in definite contexts are predicted to be available through the *iota* shift except for languages that have definite articles. This prediction follows from the type-shift ranking in (4), whereby the *iota* shift is not outranked by other type-shifts, and from the Blocking Principle in (2)—the hypothesis that overt type-shifts (e.g., definite articles) block their covert application. For indefinite contexts, the predictions differ for singular and plural BNs. Indefinite readings of singular BNs are excluded, independently of indefinite article morphology: due to the outranking of the existential type shift by the *iota* and *down* shifts (see (4)), and the number restrictions on the latter two, neither the existential shift nor DKP can derive indefinite readings of singular BNs (see (3)). As far as plural BNs are concerned, the existential type-shift is unavailable, but the *down* and *up* shifts can derive indefinite readings in tandem if the *down* shift is followed by DKP. Our understanding of DKP within the KA is that it is not a type-shift; consequently, it is insensitive to the Blocking Principle. We therefore expect no formal semantic restriction on the availability of BNs in plural indefinite contexts (but see Section 4.2.2 for a potential contribution from syntactic factors in the form of Jiang, 2020 $\pm \text{ARG}_{\text{unrestricted}}$ parameter).

For classifier languages, the KA yields the hypothesis that BNs start life as kinds (type e_k). BNs are predicted to be available in indefinite and definite contexts. Definite readings are derived through Situation Restriction (see (5)) and indefinite readings are derived through DKP. As before, we assume that DKP is not a regular type-shift and is insensitive to the Blocking Principle. All other things being equal, the prediction the KA makes for indefinite readings of BNs in classifier languages is thus that they are always available. Whether or not SR interacts with the Blocking Principle (or an adapted version thereof), is an open question we address in Section 4.4.3. For now, we assume that SR is not sensitive to the Blocking Principle and is consequently always

available according to the KA.

Language-specific predictions

To finetune the predictions for the languages in our sample, we must spell out the articles that the KA assumes for the different languages. We do so in (6):

(6) Assumptions about articles in the KA:

- **No article:** Mandarin, Hindi and Russian (Chierchia, 1998; Dayal, 2004; Jiang, 2020).
- **Definite article:** Hebrew (Doron, 2003).
- **Definite article_{SG/PL} and indefinite article_{SG}:** Spanish and German (Chierchia, 1998; Dayal, 2004; Jiang, 2020).

The list in (6) is more extensive than it would need to be since the existence of definite articles has no impact on the predictions of the KA for classifier languages based on the general predictions we worked out above. The same holds for indefinite articles in both classifier and non-classifier languages. We still opt for an exhaustive list because we will assume the descriptive adequacy of (6) to finetune the general predictions of the PA (see Section 4.3), which are sensitive to a broader range of articles. Relying on a single set of assumptions about articles seems to be the best way to come to a first balanced assessment of the KA and the PA. In Section 4.4.4, we will explore the impact that slight modifications of these assumptions have on the explanatory potential of the two approaches.

With the assumptions in (6) in place, we can finetune the general predictions made above. There are no changes for classifier languages nor for the availability of BNs in indefinite contexts in non-classifier languages. As before, Mandarin is predicted to allow for BNs in definite and indefinite contexts alike and in non-classifier languages, BNs are expected to be acceptable in plural indefinite contexts, but excluded in singular indefinite contexts. The crucial refinements concern the definite domain in non-classifier languages where the assumptions in (6) lead the KA to predict that Hindi and Russian allow for BNs whereas Hebrew, Spanish and German do not. These language-specific predictions are summarized in Table 4.1.

We mark the expected availability of BNs per context for each language in Table 4.1, using the ✓ sign where BNs are expected to be available and the × sign where BNs are expected to be unavailable.

Table 4.1: Language-specific predictions of the KA for the availability of BNs in standard argument positions in (in)definite singular/plural contexts.

	Singular		Plural	
	INDEF.	DEFINITE	INDEF.	DEFINITE
Non-classifier languages				
Spanish BSs and BPs	×	×	✓	×
German BSs and BPs	×	×	✓	×
Russian BSs and BPs	×	✓	✓	✓
Hindi BSs and BPs	×	✓	✓	✓
Hebrew BSs and BPs	×	×	✓	×
Classifier languages				
Mandarin BNs	✓	✓	✓	✓

4.3 The Properties Approach and its predictions

As we indicated in Section 4.1, the PA is cast in the same type-shifting framework as the KA, but differs from it in assuming that nouns always start life as predicates and that type-shifts are unranked with respect to one another. The PA furthermore does not allow for DKP, arguing that it involves a *sequence* of type-shifts – the *up* and the existential shifts – and consequently violates the consensus that type-shifts are last resort operations. The PA does take over the Blocking Principle as a way to restrict covert type-shifting.

The above hypotheses lead the PA to predict that every language allows its BNs to undergo the *down*, *iota* and existential shifts unless they have articles to block these type-shifts from applying covertly. The upshot is that the availability of definite and indefinite readings of BNs is fully determined by the availability of articles.

To finetune the predictions for the languages in our sample, we follow the assumptions about articles in the KA literature (see (6)). Combining these assumptions with the above predictions leads us to the language-specific predictions in Table 4.2. The table indicates that BNs are expected to be available when there is no corresponding article and it signals their unavailability exactly for those cases for which the languages do have a corresponding article: Spanish and German singular indefinite and singular/plural definites, and Hebrew singular and plural definites. The cells including an asterisk mark the differences from the predictions of the KA in Table 4.1.

Table 4.2: Language-specific predictions of the PA for the availability of BNs in standard argument positions in (in)definite singular/plural contexts

	Singular		Plural	
	INDEF.	DEF.	INDEF.	DEF.
Non-classifier languages				
Spanish BSs and BPs	×	×	✓	×
German BSs and BPs	×	×	✓	×
Russian BSs and BPs	✓*	✓	✓	✓
Hindi BSs and BPs	✓*	✓	✓	✓
Hebrew BSs and BPs	✓*	×	✓	×
Classifier languages				
Mandarin BNs	✓	✓	✓	✓

4.4 A parallel corpus study

With the predictions of the KA and the PA in place, we can compare them side by side. Table 4.1 and Table 4.2 reveal that the two approaches make very similar predictions. For plural contexts, the two approaches are indistinguishable (but see our discussion of scope in Section 4.1). For singular contexts, we do find several differences, specifically in indefinite contexts in languages for which we have assumed that they lack indefinite articles. In such contexts, the PA predicts Russian, Hindi and Hebrew BSs to appear freely, whereas the KA predicts them to be unavailable. The predicted unavailability in the KA is due to a combination of the type-shift ranking in (4) and the unavailability of the *down* shift for singulars. BNs in Mandarin escape these restrictions in the KA because they start life as non-singular kinds, allowing them to undergo DKP. Given that the predictions of the KA and PA only differ significantly for the singular domain, we will restrict our parallel corpus study to singular definite and indefinite contexts.

Before spelling out our methodology in more detail, we need to get back to the role of pseudo-incorporation that we briefly hinted at in Section 4.2. As Dayal (2004) already pointed out, the appearance of BNs in indefinite contexts may involve cases of pseudo-incorporation, which would fall outside the scope of the predictions in Table 4.1 and Table 4.2. Therefore, one should be mindful that BNs in indefinite contexts do not all necessarily appear in standard argument positions when considering the actual data. For the languages in our sample, pseudo-incorporation has been suggested to play a role in Hindi (Dayal, 2004, 2011), Spanish (Dobrovie-Sorin et al., 2006; Espinal

and McNally, 2011), Hebrew (Doron, 2003) and Mandarin (Huang, 2015; Luo, 2022). At the same time, little is known about the extent of pseudo-incorporation in (some of) these languages. The current study assumes, in line with Dayal (2011), the view that pseudo-incorporation has limited productivity in those languages that make use of it. Therefore, for our data, we will assume that the explanatory potential of pseudo-incorporation is limited to languages with a restricted distribution of BNs in indefinite contexts.

4.4.1 Methodology

As indicated in Section 4.1, we rely on a translation corpus to map out cross-linguistic variation under conditions of maximal comparability. Our corpus consists of the translations of the first chapter of *Harry Potter and the Philosopher's Stone* to Spanish, German, Hebrew, Russian, Hindi, and Mandarin. With the KA and PA leading to diverging predictions for singulars, we use the translations of *a* N_{sg} ($n = 90$) and *the* N_{sg} ($n = 140$) as arguments of verbs and prepositions as proxies for the singular definite and indefinite domains. We compare the distributions of BNs to their main competitors as they emerge from our data: the indefinite article and the numeral *one* for indefinites and the definite article and demonstrative for definites. Examples (7) and (8) illustrate the data in our corpus, presenting two English indefinite contexts with their translations to the languages of our sample.

- (7) “She told him over dinner all about Mrs. Next Door’s problems with her daughter and how Dudley had learnt **a new word** (‘Shan’t!’).”

a. **Spanish**

Mientras comían, le informó de los problemas de la señora
while were.eating.3P him informed of the problems of the Mrs
Puerta Contigua con su hija, y le contó que Dudley había
Door Next with her daughter, and her told that Dudley had
aprendido **una nueva frase** (“¡no lo haré!”).
learnt a new phrase (“no it will.do.1P!”).

b. **German**

Beim Abendessen erzählte sie ihm alles über Frau Nachbarins
at.the dinner told she him all about Mrs Neighbour’s
Probleme mit deren Tochter und dass Dudley **ein neues Wort** gelernt
problems with her daughter and that Dudley a new word learnt
hatte (“pfui”).
had (“ugh”).

c. **Hebrew**

bə-ʔaruxat ha-ʔerev hi sipra l=o ʔal kol ha-bəʔajot
 at-meal.of the-evening she told to=him about all the-problems
 še-yeš la-gveret ha-šxena ʔim ha-bat šela ve-ʔal
 that-there.is to.the-lady the-neighbor with the-daughter of.her and-about
 ze še-dadli lamad ha-jom **bituj xadaš** (lo rotse).
 it that-Dudley learnt the-day expression new (not want).

d. **Russian**

Za obedom ona oxotno spletničala, rasskazav misteru Dursley
 at lunch she gladly gossiped, having.told mister.DAT Dursley
 o tom, što u ix soседki ser'eznye problemy s dočer'ju, i
 about it, that at their neighbour serious problems with daughter, and
 naposledok soobščiv, što Dudley vyučil **novoe slovo**
 finally having.informed, that Dudley learnt new word
 “xačču!”.
 “I.wanna!”.

e. **Hindi**

Unho-ne dinner par apne pati ko bata-ya ki padosan ki apni
 She-ERG dinner on her husband to told-PFV that neighbor of own
 beti ke.sath kya samasyaye chal rahi hai aur Dudley-ne
 daughter with what problems go PROG be.PRES and Dudley-ERG
ek naya vakya sikh-a hai ‘nahi karu-nga’.
 a new sentence learn-PFV be.PRES ‘no do-FUT-M’.

f. **Mandarin**

wǎnfàn zhuō shàng, Désiǐ tàitai xiàng tā jiǎngshùle línjū jiā
 dinner table on, Dursley Mrs to he tell-ASP neighbour family
 de mǔ-nū máodùn, hái shuō Dáli yòu xuéhuì **yí-gè**
 DE mom-daughter conflict, also say Dudley again learn-RVC one-CL
xīncí (“jué-bù”).
 new-word (“never”).

- (8) “It was on the corner of the street that he noticed the first sign of something peculiar – a cat reading **a map**.”

a. **Spanish**

Al llegar a la esquina percibió el primer indicio de que
 at.the arrive at the corner noticed.3P the first sign of that
 sucedía algo raro: un gato estaba mirando **un plano**
 was.happening.3P something strange: a cat was looking a plan
de la ciudad.
 of the city.

b. **German**

An der Straßenecke fiel ihm zum ersten Mal etwas
 at the street.corner fell him for.the first time something
 Merkwürdiges auf – eine Katze, die **eine Straßenkarte** studierte.
 strange PREF – a cat, that a street.map studied.

c. **Hebrew**

rak bə-keren ha-rexov hu hivxin ba-siman ha-rifon lə-mafəhu
 only at-corner the-street he noticed in/the-sign the-first of-something
 muzar – xatula fə-ʔijna **bə-mapa**.
 weird – cat.F that-read in-map.

d. **Russian**

Tol'ko na uglu ulicy mister Dursley nakonec zametil, što
 only on corner street.GEN mister Dursley finally noticed, that
 proisxodit što-to strannoe, a zametil on košku, vnimatel'no
 happens something strange, and noticed he cat.ACC, attentively
 izučavšuju ležaščuju pered nej **kartu**.
 examining lying in.front.of her map.ACC.

e. **Hindi**

Sadak-ke mod par Dursley ko pehli ajib chiz dikh-i – ek
 street-GEN corner on Dursley to first strange thing.F see-PST.F – a
 billi, jo **naksha** padh rahi thi.
 cat.F, who map read PROG be.PST.

f. **Mandarin**

zài jiējiǎo shàng, tā kàndào le dì-yī-gè yìcháng-de
 at street.corner on, he see-RVC-ASP ORD-one-CL peculiar-DE
 xìn hào – yī-zhī-māo zài kàn **dìtú**.
 sign – one-CL-cat PROG read map.

4.4.2 Results

Figures 4.1 and 4.2 summarize the data, the former spelling out the translations of *a* N_{sg} , the latter those of *the* N_{sg} . We present a brief run-through organized per language. Spanish and German by and large come out as languages in which nouns in singular definite contexts require the definite article and nouns in singular indefinite contexts require the indefinite article. Hebrew comes out as a language in which nouns in singular definite contexts require the definite article but typically appear bare in singular indefinite contexts (Doron, 2003). The *rest* category in the Hebrew definite domain is

bigger than its Spanish and German counterparts. This is due to the frequent use of the construct state (e.g., *ʔaruxat ha-ʔerev*, ‘dinner’, in (7c)), which amounts to 44% (15/34) of the *rest* category and 12% (15/127) of all noun phrases in Hebrew definite contexts. Russian relies on BSs in definite and indefinite contexts alike (Seres and Borik, 2018). Hindi BSs have a hybrid position: they freely allow for definite readings but appear next to numeral *ek* (‘one’) N in the indefinite domain (Dayal, 2004). Mandarin BNs appear next to numeral *yi*+CL (‘one’) N in the indefinite domain (Li and Thompson, 1989) and demonstratives show an increased use in the definite domain (Jenks, 2018; Bremmers et al., 2022; Jiang and Dayal, 2023).

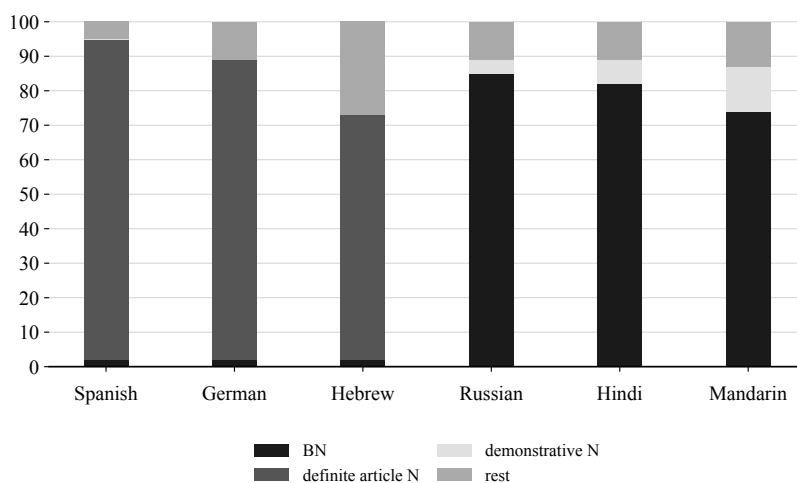


Figure 4.1: Spanish, German, Hebrew, Russian, Hindi and Mandarin translations of $a(n) + N_{sg}$ (%) in Chapter 1 of *Harry Potter and the Philosopher’s Stone*.

4.4.3 Discussion

Our data are in line with descriptive generalizations from the literature, but when juxtaposed, they reveal challenges for the KA and the PA alike. We argue that both approaches only account for part of the data.

For Spanish and German, both the KA and PA correctly predict the absence of BSs in argument position. The BSs that we do find in the indefinite domain appear after prepositions and – for Spanish – in the object position of HAVE-verbs, in line with claims about pseudo-incorporation in the literature (Dobrovie-Sorin et al., 2006; Espinal and McNally, 2011). We conclude that the KA and PA are equally successful

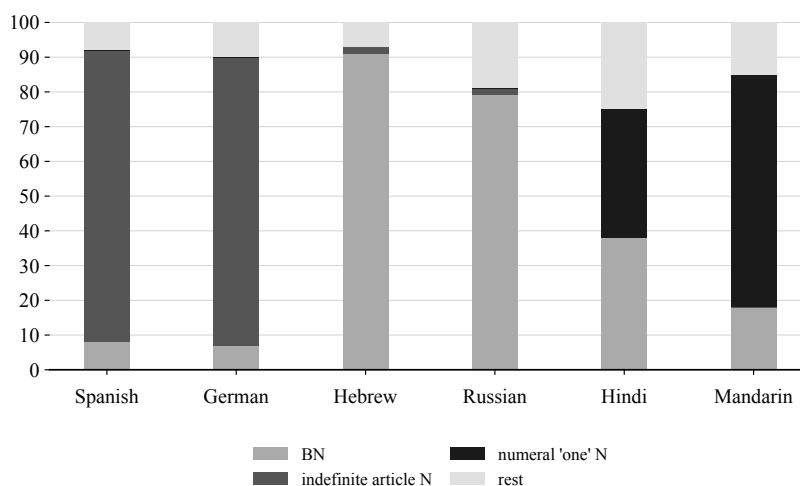


Figure 4.2: Spanish, German, Hebrew, Russian, Hindi and Mandarin translations of *the* + N_{sg} (%) in Chapter 1 of *Harry Potter and the Philosopher's Stone*.

in accounting for the Spanish and German facts.

For Hindi, we find that the PA does not make the right predictions but the KA does, *modulo* an important role for pseudo-incorporation. The Hindi data challenges the PA in the sense that the absence of definite and indefinite articles that we assumed with the literature (see (6)), leads the PA to predict BSs to freely appear in definite and indefinite contexts. The difference in distribution of BSs between these two types of contexts suggests that the PA makes the right predictions for the former but not for the latter, leaving the presence of the numeral in datapoints like (7e) as opposed to its absence in datapoints like (8e) unaccounted for. For the KA, the appearance of BSs in definite contexts is straightforwardly explained and so is the appearance of the numeral in datapoints like (7e). The KA can also account for the absence of the numeral in datapoints like (8e) under the assumption that BSs are allowed in indefinite contexts if they are pseudo-incorporated, as proposed in Dayal (2004, 2011). We submit that the opposition between (7e) and (8e) is in line with the predicted pattern. Assuming with Le Bruyn et al. (2016) that pseudo-incorporation is likelier with VO combinations in which the verb doubles a relation that is implicit or explicit in the object noun, the opposition between (7e) and (8e) follows. Indeed, *naksha* ('map') in (8e) has a READ relation as part of its telic quale (Pustejovsky, 1995) and this is doubled by the verb *padh* ('read'), predicting the availability of the BN as part of a pseudo-incorporation construction. *Vakya* ('sentence') in (7e) arguably does not come with any implicit

learning relation in its qualia, making its combination with *sikh* ('learn') unlikely to allow for pseudo-incorporation. Given our current assumptions about articles, the KA but not the PA can account for the distribution of BSs in both definite and indefinite contexts. We conclude that the Hindi data *prima facie* favor the KA over the PA.

For Hebrew and Russian, however, the tables turn, and only the PA straightforwardly makes the right predictions. The absence of definite and indefinite articles in Russian leads the PA to correctly predict BSs to freely appear in definite and indefinite contexts. The KA makes the right predictions in definite contexts but fails to extend its success to indefinite contexts, where it predicts BSs to be unavailable, contrary to fact (see (7d) and (8d)). If we were to analyze all Russian BSs in indefinite contexts as pseudo-incorporated, they would fall outside the scope of the KA's predictions for standard argument positions. However, we discard this theoretical possibility since BSs in Russian seem to lack the restricted use that pseudo-incorporation makes us expect (see Section 4.4).³ For Hebrew, the presence of a definite article and the absence of an indefinite article lead the PA to correctly predict BSs to appear freely in indefinite contexts but not in their definite counterparts. The KA, in sharp contrast to the PA's success in capturing the Hebrew data, is challenged by the availability of BSs in indefinite contexts (see (7c) and (8c)), and a pseudo-incorporation route lacks empirical support in Hebrew as it does in Russian. We conclude that the Hebrew and Russian data favor the PA over the KA.

For Mandarin, the predictions of the PA and KA are more in line with our BN data in definite than in indefinite contexts. Under the assumption that Mandarin does not have articles, the PA makes the prediction that BNs should be equally acceptable in indefinite and in definite contexts, contrary to fact. The KA faces the same problem: given that Situation Restriction and DKP are expected to be equally available for Mandarin BNs, the KA fails to predict the marked difference we find in their distribution between definite and indefinite contexts. For both approaches, the relatively high proportion of demonstratives in definite contexts (13%) – especially in comparison to Russian (4%) – also comes as a surprise. We conclude that the Mandarin data are problematic for both the KA and the PA.

Taking stock, we have argued that the KA and PA make the right predictions for Spanish and German. For Hebrew, Russian and Hindi, we have argued that both approaches make the right predictions for BNs in definite contexts but only successfully account for BNs in indefinite contexts for a subset of the languages: as it stands, the PA makes the right predictions for Hebrew and Russian but not for Hindi, whereas

³See Mueller-Reichau (2015) for a detailed discussion of pseudo-incorporation in Russian.

the KA makes the right predictions for Hindi but not for Hebrew or Russian. Finally, for Mandarin, we find that both approaches have trouble accounting for BNs in both indefinite and definite contexts.

4.4.4 The explanatory potential of the PA and the KA

With the predictions of neither approach being fully borne out, the question that imposes itself is whether we can tweak either or both to reach empirical adequacy. We start with definite contexts and then move to indefinite ones.

For definite contexts, the only language that leads to problems for the PA and KA is Mandarin. The recent literature is converging on the idea that Mandarin BNs and demonstratives are in complementary distribution even though the details remain to be worked out (see, e.g., Jenks, 2018; Bremmers et al., 2022; Simpson and Wu, 2022). We are confident that both approaches can be extended to cover the data. For the PA, this would involve a refinement of the Blocking Principle. For the KA, something akin to the Blocking Principle would need to be developed to model the interaction between the demonstrative and BNs at the level of SR.

For the indefinite data, we first discuss possible extensions of the PA and then move to the KA. We argue that the PA can straightforwardly be extended to cover the totality of the data. To do so, our first step is to change our assumptions about the cross-linguistic inventory of indefinite articles: we originally followed the KA literature in assuming that Hebrew, Russian, Hindi, and Mandarin lack indefinite articles. Our proposal is to change this assumption for Hindi and Mandarin and to take Hindi *ek* and Mandarin *yi+CL* to function as indefinite articles. With this assumption in place, the PA predicts the Blocking Principle to kick in and block Hindi BSs and Mandarin BNs from appearing in singular indefinite contexts.

Next, we assume with Dayal (2004, 2011) that the availability of Hindi BSs in singular indefinite contexts relies on pseudo-incorporation and we extend this assumption to Mandarin BNs (see also Huang, 2015; Luo, 2022). It follows from this position that the PA no longer predicts Hindi BSs and Mandarin BNs to be excluded from singular indefinite contexts, but rather that their use should be restricted. This prediction is in line with the tendencies we find in the frequency data in Figures 4.1 and 4.2. It also straightforwardly explains why the presence/absence of the Hindi numeral in (7e) and (8e) neatly patterns with the presence/absence of the Mandarin numeral in (7f) and (8f).

Moving to the KA, we argue that the extensions we proposed for the PA do not

affect the predictions of the KA and explain why we think that the latter cannot be extended to cover the totality of the data. As for the extensions we proposed for the PA, the assumption that Hindi *ek* functions as an indefinite article merely doubles the restriction in the KA on BSs in singular indefinite contexts that follows from the ranking of the *iota* shift above the existential shift. The assumption that Mandarin *yi+CL* also functions as an indefinite article furthermore has no impact on the predictions of the KA given that DKP is not part of the regular set of type-shifts and the Blocking Principle does not apply to it. We conclude that the extensions we proposed for the PA do not have a direct impact on the predictions of the KA.

As for the difficulties we see to extend the KA to cover the totality of the data, we think it would be feasible to cover the Hebrew data but neither the Russian nor the Mandarin data. For Hebrew, one could work out an analysis in which the existential type-shift becomes available to BSs because the *iota* and the *down* shifts are independently blocked. A similar escape route is not available in Russian, where the availability of BSs in definite contexts shows that the *iota* shift is not blocked. To cover the unrestricted availability of Russian BSs in indefinite contexts, the only path we see for the KA is to abandon the assumption that the *iota* shift is ranked above the existential shift. However, this would mean that a fundamental insight of the approach must be abandoned. As for the restriction we find on the use of BNs in indefinite contexts in Mandarin, we see no easy way to account for it unless we assume DKP is simply not available, a yet more problematic move than reconsidering the type-shift ranking. The problem is that Mandarin BNs are unrestricted in plural indefinite contexts (Liu et al., 2023b), strongly suggesting that the restrictions in singular indefinite contexts really come from an interaction between BNs and *yi+CL*. However, the KA—in its current version—has no level at which DKP could be made to interact with the regular type-shifting operations that are involved in the use of the numeral.

Summarizing the discussion, we have argued that both the PA and the KA are likely to be able to account for the use of BNs in definite contexts but that only the PA can easily be extended so as to cover the full set of indefinite data we found. The main challenges for the KA lie in (i) the clearcut opposition between Russian and Hindi BNs in indefinite contexts and (ii) the restricted distribution of BNs in Mandarin. For the PA, we have argued that it can be extended to cover the totality of the data if we assume—contrary to the consensus in the KA literature—that the numeral *one* in Hindi and Mandarin functions as an indefinite article and if we furthermore assume that pseudo-incorporation can account for the indefinite uses of BNs in these languages in singular indefinite contexts. Obviously, the latter assumption calls for a follow-up

study in which BNs in indefinite contexts are studied through the lens of pseudo-incorporation, paying close attention to the diverging distribution of BSs/BNs in the languages for which we take pseudo-incorporation to be at play (Spanish, Hindi, and Mandarin).

4.4.5 Two notes on methodology

We end our discussion with two methodological notes regarding parallel corpus research in general and the level of parallelism we have pursued in the current study. Le Bruyn and de Swart (2024) take parallel corpus research to be a valuable addition to the cross-linguistic semanticist's toolbox, but one that ultimately requires replication and triangulation to accumulate critical mass across studies. The current paper is best considered in this spirit as a proof of concept. As such, the data it brings to bear on the availability of BNs should be taken with a grain of salt, and the conclusions that we draw from it handled with care. For future work, we intend to run the same analysis on the same HP chapter, translated by different translators for a subset of the languages in our sample, namely, Russian, Hindi, and Mandarin, which all lead to diverging predictions for the PA and the KA. For these languages, the first HP volume happens to have at least two official translations, allowing us to assess how representative our data is for these languages. We also wish to extend the number of languages we examine and refer the interested reader to Borik et al. (2025) for a closer look at parallel data from Polish and Macedonian, the former patterning with Russian, the latter patterning with Hebrew in the definite domain but exhibiting a more extensive use of the numeral *one* in the indefinite domain.

Regarding the level of parallelism we have pursued in this comparative study, the attentive reader may have noticed that we present our frequency data per expression per language but not per context. The upshot of this is that one cannot directly evaluate how the contexts in which an expression α in one language appears relate to the contexts in which an expression β in another language appears. The choice not to go for parallelism at the level of contexts is inspired by the size of our corpus: every translation brings in a limited number of idiosyncratic choices, but these add up with every language we add, making it hard to discern the bigger patterns in our fairly small dataset if we present the data per expression per language per context. In future work, we will be adding more chapters of the same translations and pursue parallelism at the level of contexts in the way we have done in other studies under the *Translation Mining* approach (see, e.g., Bremmers et al., 2022; van der Klis et al., 2022). This

will allow us—in particular—to really probe the variation in pseudo-incorporation that emerges from our data.

4.5 Conclusion

In this paper, we adopted a translation corpus approach to reassess the empirical coverage of two closely related theories of argument formation: Chierchia’s Kinds Approach (KA) (Chierchia, 1998; Dayal, 2004; Jiang, 2020) and Krifka’s (2003) Properties Approach (PA). Given that both theories make the same predictions for plural contexts, we focused on singular contexts. Our corpus consisted of the translations of the first chapter of *Harry Potter and the Philosopher’s Stone* to Spanish, German, Russian, Mandarin, Hindi, and Hebrew.

We argued that both the KA and the PA make the right predictions for singular definite contexts but that only the PA can be extended to account for the patterns we found for singular indefinite contexts. To derive these patterns in the PA, we hypothesized that the numeral *one* in Hindi and Mandarin functions as an indefinite article and that the BNs that appear in singular indefinite contexts in the two languages are to be accounted for using pseudo-incorporation. The challenges we identified for the KA lie in the unconstrained distribution of BNs in singular indefinite contexts in Russian and in the constrained distribution of BNs in the same contexts in Mandarin. The first challenge calls the KA’s type-shift ranking into question, the second the status of DKP.

Relevant follow-up research that we identified includes replication of the findings on the basis of a second set of official translations of the same corpus for Russian, Hindi and Mandarin as well as an extension of the current corpus to pursue parallelism at the level of contexts and probe the variation we find in more detail. Special attention would need to be paid to the varying extent of pseudo-incorporation, especially in Spanish, Hindi and Mandarin.

CHAPTER 5

Fine-Tuning the Property-Based Analysis: the Indefinite Domain

5.1 Introduction

The preceding chapters have followed a path of empirical discovery, moving from language-specific to cross-linguistic validated patterns that challenge existing formal semantic theories of argument formation. We have established in Chapters 2, 3, and 4 a crucial empirical pattern for the Mandarin indefinite domain: the systematic alternation between the numeral-*yi* construction with high frequency and bare nouns with distributional restrictions. The corpus study in Chapter 2 compared the referential system of English and Mandarin. From the alternation between numeral-*yi* and bare nouns in indefinite contexts emerges numeral-*yi* as the dominant form. The cross-linguistic comparison in Chapter 3 confirmed that this pattern is not a corpus bias but a genuine feature of Mandarin; a similar pattern was not replicated in Russian or Hindi translations, two other languages without articles. Chapter 4 further validated the reliability of the translation data by testing meaning stability across a broader typological range. The extension to a multilingual comparison with languages with full article systems (Spanish, German) and one with only a definite article (Hebrew) demonstrated that translation patterns in these languages align with their known grammatical rules

for (in)definiteness. The cross-linguistic patterns constituted the empirical basis for the evaluation of the explanatory power of the Kinds Approach and the Properties Approach. While the Kinds Approach struggled to account for the variation across Mandarin, Russian, and Hindi, the Properties Approach has the flexibility to adapt to the observed pattern. Specifically, we proposed two core modifications to the Properties Approach, as hypotheses to resolve the alternation challenge for Mandarin pattern of bare nouns and numeral-*yi* in indefinite contexts:

- (i) Numeral-*yi* functions as an indefinite article and blocks bare nouns from appearing in regular, indefinite argument positions.
- (ii) Bare nouns in singular indefinite contexts are predicative and restricted to pseudo-incorporation constructions.

Together, these hypotheses explain the alternation between numeral-*yi* and bare nouns: while numeral-*yi* marks regular indefinite arguments, bare nouns survive only in pseudo-incorporated positions. This division of labor resolves the alternation challenge within the Properties Approach by showing that the two forms do not actually compete for the same grammatical space. Because numeral-*yi* functions as the default indefinite article, it blocks bare nouns from regular argument positions—as predicted by the Blocking Principle—and relegates them to non-argumental, pseudo-incorporated positions through a separate mechanism. Bare nouns escape the blocking effect because they are not true arguments competing with the overt indefinite article.

The core of this proposed division of labor between numeral-*yi* and bare nouns hinges on the premise that the Mandarin bare nouns in indefinite contexts are pseudo-incorporated. However, this remains a theoretical proposal. The present chapter undertakes the next step with one overarching goal: to empirically test and theoretically formalize the pseudo-incorporation analysis of Mandarin bare nouns in indefinite contexts. Achieving this goal requires us to overcome a significant empirical obstacle: Mandarin lacks the clear morpho-syntactic hallmarks (e.g., special case or number marking) typically used to identify pseudo-incorporation cross-linguistically (Borik and Gehrke, 2015). To meet this challenge, we adopt and extend a lexical-semantic criterion from Huang (2015): the typicality of the relationship between a verb and its object. We hypothesize that if Mandarin bare nouns in indefinite contexts are pseudo-incorporated, they should be restricted to combinations with verbs to which they bear a stereotypical or institutionalized relation (e.g., read book, drive car), rather than incidental ones (e.g., find book, wash car).

We test this prediction through a new, targeted translation corpus study based on the full *Harry Potter and the Philosopher's Stone* novel, particularly focusing on Mandarin translations of English NPs preceded by *alan* in the object positions. If our hypothesis on pseudo-incorporation is correct, we can distinguish the division of labor between bare nouns and numeral-*yi* in the indefinite domain in the way we proposed for the modified Properties Approach: bare nouns appear almost exclusively in verb-object combinations with typicality as pseudo-incorporation, and numeral-*yi* functions as the default option in all other indefinite contexts.

The structure of this chapter is as follows. In Section 5.2, we position our account within the existing literature on the articlehood of numeral-*yi* and Mandarin pseudo-incorporation, clarifying what is new in our proposal. Sections 5.3 to 5.5 are dedicated to the empirical investigation: we review diagnostics for pseudo-incorporation, present our corpus study, and demonstrate that the distribution of bare nouns is indeed governed by the typicality restriction. Finally, in Section 5.6, we develop a novel formal analysis within the Properties Approach that captures this restriction and positions Mandarin pseudo-incorporation within a broader typology. Section 5.7 concludes the chapter.

5.2 On articles and pseudo-incorporation in Mandarin

The hypotheses we proposed in Chapter 4 are not new. The functionalist literature has argued that numeral-*yi* (henceforth, *yi*) is on a grammaticalization path to articlehood (see Chapter 1) and Mandarin bare nouns have been argued to appear in pseudo-incorporation constructions (Lü, 1955; Chao, 1968; Lü, 1979; Li and Thompson, 1989; Zhu, 1982; Barrie and Li, 2015; Huang, 2015; Luo, 2022). What is new, though, is that we take these hypotheses to jointly account for the distribution of bare nouns in singular indefinite contexts in Mandarin. The account we pursue is that *yi* is the overt realization of the existential type-shift and blocks bare nouns from appearing in singular indefinite argument positions, limiting their occurrence to pseudo-incorporation constructions. In what follows, we give a brief sketch of the literature on the article status of *yi* and the literature on pseudo-incorporation. This will allow us to define how the account we pursue differs from what is typically assumed in these literatures and to identify the challenges that lie ahead.

The article status of *yi*

In the account we pursue, we take *yi* to be an indefinite article and to block bare nouns from appearing in singular indefinite argument positions. In the functionalist literature, the role of *yi* as an article is rarely framed in terms of (total) blocking. Rather, this literature appeals to pragmatic factors like referential importance as favoring the use of *yi* (Wright and Givón, 1987). One quantitative measure for referential importance is the average recurrence rate of referents: whereas referents introduced by bare nouns recur 1.87 times in Wright & Givón's corpus, those introduced by *yi* recur 14.5 times.

We submit that a type-shifting account of *yi* is not a major departure from the analysis of *yi* in the functionalist literature. What is clear, though, is that our account will need to address the differences in referential importance between bare nouns and *yi* arguments. We do not consider this a major challenge and postpone doing so until later in this chapter (Section 5.6.5). The gist of our proposal will be that the reduced recurrence rate of referents of bare nouns has to be related to the reduced discourse transparency of pseudo-incorporated nouns (see, e.g., Farkas and Swart, 2003 on Hungarian and Dayal, 2003, 2011 on Hindi).

Mandarin pseudo-incorporation

The account we pursue takes all bare nouns in singular indefinite contexts to be pseudo-incorporated. The literature on Mandarin pseudo-incorporation has a less liberal take on its range and typically focuses on examples like the ones in (1) and (2), taken from Huang (2015):

- (1) kàn shū
read book
'read a book'
- (2) zuò mèng
make dream
'dream a dream'

According to Huang (2015), Mandarin cases of pseudo-incorporation can be characterized as verb-noun combinations that involve a relation of typicality between the verb and the noun. Based on Huang's examples, we interpret this typicality relation as one in which the verb describes a typical use of the object corresponding to the noun, as in (1), or the typical action that brings the object corresponding to the noun into existence, as in (2). Luo (2022) adds that these verb-noun combinations come

with a number of properties that have been linked to pseudo-incorporation in other languages. Next to reduced discourse transparency (cf. *supra*), these include narrow scope and restricted modification:

- (3) [zuó wǎn] tā méiyǒu kàn **shū**
 last night he NEG read book
 ok ‘He didn’t read any book.’
 # ‘There’s **a book** he didn’t read.’
- (4) [zuó wǎn] tā méiyǒu kàn **yì běn shū**
 last night he NEG read one CL book
 ok ‘There’s **a book** he didn’t read.’
- (5) wǒ zhèngzài kàn ***(yì běn) hǎo shū**
 I ASP read one CL good book
 ‘I’m reading **a good book**.’

The restricted set of examples discussed in the literature on pseudo-incorporation suggests that an extension of a pseudo-incorporation analysis to all bare nouns in indefinite singular contexts is not straightforward. However, it is important to distinguish between Huang’s typicality criterion and the other semantic criteria discussed by Luo. Luo notes that the criteria he proposes do not uniformly apply to his curated set of cases of pseudo-incorporation. It is consequently clear that they would *a fortiori* not apply to all bare nouns. Huang’s typicality criterion is different in that it has never been looked into for other verb-noun combinations than the typical set we find in the literature on pseudo-incorporation. It is thus an open question whether the typicality criterion applies to all cases of verbs combining with bare nouns.

We did not see major challenges in the literature on the article status of *yi* but we do see significant challenges in the literature on pseudo-incorporation. The first is that not all bare nouns in Mandarin have the semantic properties that are typically associated with pseudo-incorporated nouns. Luo already signals this for his curated set of pseudo-incorporation cases and this raises the question whether a generalized pseudo-incorporation analysis of bare nouns is feasible at all. We tackle this challenge in Section 5.3. There, we argue that the semantic properties of pseudo-incorporated nouns are unstable across and within languages and that the non-uniformity of semantic properties like discourse transparency and restricted modification for Mandarin bare nouns does not endanger a pseudo-incorporation analysis.

In Section 5.4, we move to our second challenge. If we succeed in arguing that Mandarin bare nouns do not need to show all the properties that have been associated

with pseudo-incorporation in other languages, there should still be a stable property that allows us to maintain that Mandarin bare nouns are always pseudo-incorporated. The key, we argue, lies in moving from properties of bare nouns to properties of verb-noun combinations. We make this move on the basis of the cross-linguistic literature on pseudo-incorporation and argue that Huang's typicality criterion aligns with restrictions that have been noted for verb-noun combinations in languages like Hindi. In Section 5.5, we will put the criterion to the test and establish that it successfully predicts which contexts allow for and which contexts proscribe the use of bare nouns. This result provides a clear empirical argument in favor of a pseudo-incorporation analysis of bare nouns.

In Section 5.6, we move to our third and final challenge, *viz.* to build a pseudo-incorporation analysis that derives the role of the typicality criterion and has the analytical power to position Mandarin in a broader typology of pseudo-incorporation.

With heavy challenges ahead, it is good to conclude this section by clarifying what the successful tackling of the challenges buys us: grip on a subdomain of referentiality that has eluded formal analyses up till now. In the formal literature, the only restriction that we know of for bare nouns in indefinite contexts is that they are restricted to the object position (Cheng and Sybesma, 1999). From this perspective, the extremely reduced distribution of bare nouns in indefinite contexts that we found in Chapter 2 and whose linguistic relevance was confirmed in Chapters 3 and 4 does not make sense. By adopting a blocking analysis of *yi*, our account derives that all bare nouns in indefinite singular contexts have to be pseudo-incorporated and paves the way for an account of their restricted distribution.

5.3 Diagnosing pseudo-incorporation: Part I

In this section and the next, we will review pseudo-incorporation diagnostics. We do so with two of the challenges we identified in Section 5.2 in mind. The first challenge was that a generalized pseudo-incorporation analysis of bare nouns in indefinite singular contexts in Mandarin could be a non-starter if Mandarin bare nouns were exceptional in not all aligning with the semantic properties that the literature associates with pseudo-incorporated nouns. If we can successfully overcome this first challenge, the second one presents itself, *viz.* the need to identify a property that all Mandarin bare nouns share and that links them to pseudo-incorporation. The first half of this section focuses on the semantic properties of bare nouns. In the second half, we move to their morpho-syntactic properties. After formulating an interim conclusion on diag-

nostics at the end of this section, we turn to the properties of verb-noun combinations in Section 5.4, highlighting the relevance of Huang’s typicality criterion in diagnosing pseudo-incorporation.

Semantic properties of pseudo-incorporated nouns

The literature frequently refers to a cluster of semantic properties that are associated with pseudo-incorporated nominals: (i) obligatory narrow scope, (ii) reduced discourse transparency, (iii) restricted modification possibilities, and (iv) number neutrality (see, e.g., Borik and Gehrke, 2015)¹. We discussed the first three in Section 5.2 for Mandarin. Here, we illustrate property (iv) on the basis of the Hungarian contrast in (6) and (7), taken from Farkas and Swart (2003).

(6) #Mari gyűjt **egy bélyeget**.
 Mari collect a stamp.acc

(7) Mari **bélyeget** gyűjt.
 Mari stamp.acc collect
 ‘Mari is collecting stamps.’

Farkas and Swart (2003) argue that the standard object position in Hungarian is to the right of the verb and typically requires singular nouns to be preceded by a determiner as in (6). The infelicity of this example is semantic and is due to the fact that the verb *gyűjt* (‘collect’) requires a plural object while *egy bélyeget* (‘a stamp’) is singular. Example (7) is different in that the object appears to the left of the verb, a position that Farkas & de Swart associate with (pseudo-)incorporation. Hungarian pseudo-incorporated nominals keep their case marking but are typically number neutral, explaining why (7) is felicitous, despite the fact that it contains the same singular noun *bélyeget* (‘stamp’) as (6).

Even though properties (i) to (iv) are often used to argue in favor of the pseudo-incorporated status of bare nouns, only property (i) is arguably a necessary condition. As discussed in Le Bruyn et al. (2016), properties (ii) to (iv) are not stable across and sometimes not even within languages. A comparison of (8) and (9) is indicative in this respect (from Espinal and McNally, 2011 and Lazaridou-Chatzigoga, 2011):

¹Borik and Gehrke (2015) combine the discussion of restricted modification possibilities with the discussion of well-establishedness. We prefer to separate these and get back to the latter in our discussion of combinatorial restrictions in Section 5.4.

- (8) Avui porta **faldilla**_i. #**La**_i hi vam regular l'any passat.
 today wears skirt. it her have given the year last
 'Today she's wearing **a skirt**. We gave **it** to her as a present last year.'
- (9) Foruse **pukamiso**_i htes. **To**_i ihe aghorasi sti varkeloni.
 wore shirt yesterday. it had bought in-the Barcelona
 'Yesterday he had **a shirt** on. He had bought **it** in Barcelona.'

Example (8) is from Catalan and illustrates how the bare noun *faldilla* cannot be felicitously picked up by the pronoun *la* ('it'). We assume with Espinal & McNally that singular bare nouns in Catalan are pseudo-incorporated and this example thus illustrates the reduced discourse transparency of Catalan pseudo-incorporated nominals. Lazaridou-Chatzigoga argues that singular bare nouns in Greek are also pseudo-incorporated. She indicates that the exact counterpart of (8) is equally infelicitous in Greek but adds that the minimal variant in (9) is perfectly acceptable. This comparison shows that property (ii) is not stable across or within languages. Le Bruyn et al. provide similar comparisons for properties (iii) and (iv), leading us to conclude that properties (ii) to (iv) are best interpreted in terms of family resemblance and not as necessary properties of pseudo-incorporation. As for property (i)—the obligatory narrow scope behavior of pseudo-incorporated nouns—we note that it is a necessary but not a sufficient condition, bare nouns across languages often leading to a narrow scope reading, independently of their pseudo-incorporated status (see Le Bruyn and de Swart, 2022 for recent discussion).

Morpho-syntactic properties of bare nouns

The preceding discussion of the semantic properties of pseudo-incorporated nouns shows that they can definitely be used to strengthen an argumentation in favor of the pseudo-incorporated status of nouns. However, by themselves, they do not constitute knock-down arguments in favor or against the identification of a noun as being pseudo-incorporated. It should not come as a surprise then that most cases of pseudo-incorporated nouns discussed in the literature typically exploit their morpho-syntactic properties, arguing that the nouns under scrutiny behave differently from what is considered 'standard' in the language. E.g., the appeal of a pseudo-incorporation analysis for a noun like *bélyeget* in (7) comes from the fact it occurs to the left of the verb, going against the standard VO word order in Hungarian. Form and meaning going hand in hand in formal semantics, the exceptional status of *bélyeget* at the level of its

morpho-syntax constituted a knock-down argument for Farkas and Swart (2003) to explore an equally exceptional analysis at the level of its semantics. Next to word order, the literature on pseudo-incorporation has relied on the exceptional absence/presence of determiners, case and number marking to successfully argue for the existence of pseudo-incorporation across a variety of languages (see, e.g., Espinal and McNally, 2011 and Dayal, 2011).

Taking stock

We have argued that pseudo-incorporated nominals show a cluster of semantic properties, none of them however constituting a foolproof diagnostic for pseudo-incorporation. Successful accounts of pseudo-incorporated nominals have consequently always argued for pseudo-incorporation based first and foremost on the basis of morpho-syntactic properties. We conclude that variation in the semantic properties of Mandarin bare nouns does not constitute an argument against their pseudo-incorporated status. With this, we have overcome the first challenge we identified in Section 5.2. The second challenge was to define a property that all Mandarin bare nouns do share and that links them to pseudo-incorporation. We have not met this challenge yet as it should be clear that the morpho-syntactic properties of Mandarin bare nouns will not be of much use to argue in favor of their pseudo-incorporated status: Mandarin is a rigid VO language with no overt marking of case or number. One might argue that our data do establish that the ‘standard’ use of nouns in indefinite singular contexts is with *yi* and that this allows us to argue that the use of bare nouns requires a special treatment. We recall, though, that our aim is to make the bareness of bare nouns follow from their pseudo-incorporated status, requiring us to provide independent evidence for the latter. In Section 5.4, we argue that Huang’s typicality criterion for verb-noun combinations fits the bill.

5.4 Diagnosing pseudo-incorporation: Part II

In Section 5.3, we argued that the semantic properties of pseudo-incorporated nominals should be analyzed in terms of family resemblance. This means that variation in the semantic properties of Mandarin bare nouns in singular indefinite contexts does not *a priori* exclude a pseudo-incorporation analysis. We also looked into the morpho-syntactic properties of pseudo-incorporation, arguing that these typically provide a stronger case than semantic properties while at the same time acknowledging that

they are of no help to us for diagnosing pseudo-incorporation in Mandarin. The ultimate conclusion then is that there is no knock-down argument against or in favor of a pseudo-incorporation analysis of Mandarin bare nouns in indefinite singular contexts if we limit ourselves to the semantic and morpho-syntactic properties of the nouns themselves. In this section, we move from the properties of nouns to restrictions on verbs and verb-noun combinations as an extra diagnostic for pseudo-incorporation in a number of languages. At the end of this section, we argue that Huang's typicality criterion aligns with restrictions on verb-noun combinations that have been observed for pseudo-incorporation in languages like Hindi. In Section 5.5, we put this criterion to work and check whether it allows us to predict which verb-noun combinations allow for or proscribe bare nouns.

Farkas and Swart (2003) treat pseudo-incorporation in Hungarian as a fully productive morpho-syntactic process for which they provide a semantics that imposes no restrictions on verbs, nouns or combinations thereof. As such, Hungarian entered the formal semantics literature as a language with unrestricted pseudo-incorporation (see, however, Kiefer (1990) and Kiefer and Németh (2019) for an interaction with aspect). Hungarian is not the only language in which pseudo-incorporation does not come with restrictions (see, e.g., Massam, 2001, 2020 on Niuean), but other languages have been argued to come with restrictions on the verbs allowing for pseudo-incorporation or on verb-noun combinations. Spanish, e.g., has been argued to put restrictions on the types of verbs (inspired by Espinal, 2010):

- (10) El hombre llevaba **traje**.
 the man wore suit
 'The man wore **a suit**.'
- (11) *Limpio **traje**.
 clean suit
 Intended: 'I'm cleaning **a suit**.'

Assuming that, as for Catalan (see our discussion of (8)), singular bare nouns in Spanish are pseudo-incorporated, the acceptability of *traje* ('suit') after *llevaba* ('wore') and its unacceptability after *limpio* ('I'm cleaning') suggest that pseudo-incorporation in Spanish is restricted to a subclass of verbs (see, e.g., Dobrovie-Sorin et al., 2006; Espinal, 2010; Espinal and McNally, 2011), the empirical generalization being that verbs akin to have are the ones that allow for pseudo-incorporation. The same contrasts with similar clusters of verbs have been noted for—among others—Catalan (e.g., Espinal and McNally, 2011), Norwegian (Borthen, 2003), Greek (e.g.,

Alexopoulou and Folli, 2011; Lazaridou-Chatzigoga, 2011; Gehrke and Lekakou, 2013) and Romanian (Dobrovie-Sorin et al., 2006).

Taking the step from restrictions on verbs to restrictions on verb-noun combinations, Dayal argues that there are restrictions on pseudo-incorporation in Hindi and that these are to be situated at the level of verb-noun combinations and not at the level of verbs. The examples in (12) to (14) present the relevant paradigm (from Dayal, 2011).

- (12) **laRkii**-dekhnaa
girl-seeing
- (13) ***aurat**-dekhnaa
woman-seeing
- (14) ***laRkii**-sulaanaa
girl-putting-to-sleep

Assuming with Dayal that Hindi singular bare nouns are pseudo-incorporated, the contrast between (12) and (13) shows that the same verb can lead to acceptable and unacceptable cases of pseudo-incorporation. The contrast between (13) and (14) shows that – *mutatis mutandis* – the same holds for nouns. Taken together, the paradigm in (12) to (14) shows that Hindi pseudo-incorporation is not fully productive and that the relevant restrictions are to be situated at the level of verb-noun combinations. Dayal’s empirical generalization is that the verb-noun combinations that allow for pseudo-incorporation in Hindi are those that refer to an institutionalized activity or state. (12) qualifies in the sense that it is used to refer to events in which someone is looking for a wife for her son – a significant responsibility of mothers in India. Dayal assumes that no such institutionalized activity can be linked to (13) or (14).

In this section, we have seen how the literature on pseudo-incorporation has not only looked at the properties of pseudo-incorporated nouns but also at those of verbs and verb-noun combinations. The empirical picture in the literature points at three types of languages: (i) those with no requirements on verbs, nouns or combinations thereof (e.g., Hungarian), (ii) those with lexical requirements on verbs (e.g., Spanish), and (iii) those with semantic requirements on verb-noun combinations (e.g., Hindi). Based on the initial characterization of Mandarin pseudo-incorporation by Huang as involving a typicality relation between the verb and the noun (cf. Section 5.2), we argue that Mandarin is similar to Hindi in requiring verb-noun combinations that refer to an institutionalized activity or state and that it qualifies as a language of type (iii).

With semantic and morpho-syntactic properties of pseudo-incorporated nouns providing us with insufficient grip on Mandarin pseudo-incorporation, we will consequently build our case for the pseudo-incorporated status of bare nouns in indefinite singular contexts in Mandarin on the basis of the typicality criterion. We will do so on the basis of a corpus study in Section 5.5. If successful, we overcome the second challenge identified in Section 5.2 and can move ahead to the third and final one, *viz.*, developing an analysis that derives the typicality restriction on verb-noun combinations in Mandarin with the analytical power to position Mandarin pseudo-incorporation in the broader typology sketched above.

5.5 Mandarin pseudo-incorporation

In Section 5.4, we hypothesized that Mandarin pseudo-incorporation is subject to the requirement that the verb and noun involved stand in a typicality relation to one another. In this section, we build on this hypothesis and ask whether all bare nouns in indefinite singular contexts appear in verb-noun combinations that respect this typicality restriction. If they do, we provide empirical support for a generalized pseudo-incorporation analysis of Mandarin bare nouns in indefinite singular contexts and overcome the second challenge we set for ourselves in Section 5.2. To check whether all bare nouns in indefinite singular contexts appear in verb-noun combinations that respect the typicality restriction, we set up a study that targets all indefinite singular contexts in a corpus. If the typicality restriction is key, we predict bare nouns to only appear in those contexts in which the typicality restriction is met. In all other contexts, we predict *yi + N* to be the only option.

Corpus

The corpus we use consists of all chapters of the Mandarin translation of *Harry Potter and the Philosopher's Stone*. We selected this corpus for two reasons. The first is that it is a straightforward extension of the corpus we used in earlier chapters. The second is that a Mandarin translation of a text written in a language like English provides us with an easy way to target singular indefinite contexts.

Data Collection

We automatically extracted all instances of *a(n) + N* from the English original ($n = 848$). We manually further restricted this set to those instances occurring in the object

position of verbs ($n = 331$). In the next step, we aligned the Mandarin translations to their English counterparts and filtered out cases in which $a(n) + N$ had no translation, was not translated by $yi + N$ or a bare noun, did not appear in regular object position or occurred as part of the definite-like *ba* construction. The final dataset we obtained consists of 154 nominal expressions, 31 bare nouns and 133 cases of $yi + N$. The ratio of bare nouns and $yi + N$ in this dataset is virtually identical to the ratio we found for the dataset of Chapter 4.

Annotation

We annotated every datapoint for the variables in a) to f). Next to each variable, we indicate the possible values and corresponding frequencies.

a)	form	2 values: bare noun <i>yi</i> ('one')	(31 133)
b)	typicality	2 values: yes no	(50 104)
c)	referential	2 values: yes no	(114 41)
d)	modifier	2 values: yes no	(89 65)
e)	presence of <i>le</i>	2 values: yes no	(33 121)
f)	NEG/DIS	2 values: yes no	(20 134)

Annotations a) and b) are the core annotations of interest. a) tracks the form and specifies whether the nominal expression is a bare noun or $yi + N$. b) is a judgement of whether a verb-noun combination counts as respecting the typicality restriction. The criterion we applied here was whether the verb refers to a typical use of the noun or refers to an action that brings the referent of the noun into existence. (15) to (18) illustrate the decisions we made:

(15) typicality **yes**

yífù yímā zhèngzài kàn **jìngcāi diànshì jiémù**
uncle aunt ASP watch quiz TV programme

'His uncle and aunt were watching **a TV show**.'

(16) typicality **yes**

Hèmǐn Gélánjié duì yī wèi lǎoshī sāxià le **mítian dà huǎng**
Hermione Granger to one CL teacher tell ASP enormous big lie

'Hermione Granger told **a whopper** to a teacher?'

- (17) typicality **no**
 [context: ‘Oh I’ve been reading’]

Hǎigé shuō zhe cóng tā de zhěntóu dǐxià chōu chū yì běn dà
 Hagrid say ASP from him DE pillow underneath pull out one CL big
bùtóu de shū.
 volume DE book

‘...said Hagrid, pulling **a large book** from under his pillow.’

- (18) typicality **no**

Hǎigé chěkāi zhǐbāo kànjiàn yí jiàn hòuhòu de xiānlǜsè de
 Harry opened package see one CL thick DE light.green DE
shǒu-biān máoyī
 hand-knitted sweater

‘Harry opened the package and saw **a thick, bright green hand-knitted sweater.**’

For (15), we assume that *TV shows* are made to be watched and for (16), we assume *lies* come into existence by telling them. (17) and (18) are different in the sense that the verb in neither of them is reporting on how *books* or *sweaters* come into existence. *Books* are also not made to be pulled from under one’s pillow and *sweaters* are not made to be seen.

We added annotations c) to f) to control for a number of other factors that were frequent enough and that we assumed could have an influence on the choice between a bare noun and *yi* + N. c) controls for the influence of referentiality, **no** applying to bare nouns and cases of *yi* + N occurring in the scope of modal operators or negation, **yes** applying to all other cases. d) tracks the influence of the presence of modifiers and e) the presence of the aspect marker *le*. f) requires a bit more explanation. NEG/DIS applies to nominal expressions that appear in the scope of negation or a distributivity operator. The reason we combined these is that the scope of negation and that of distributivity operators count as contexts in which singular and plural interpretations cannot be distinguished:

- (19) They were each given a match/matches.
 (20) Most of them had never seen an owl/owls even at nighttime.

In (19), the singular and plural versions of *match* are compatible with each receiving only a single match. In (20), the preferred reading of both the singular and the plural versions is that not a single owl was seen by anyone. On an article analysis of *yi*, it can only be taken to block the existential type-shift in singular indefinite contexts and not in plural indefinite contexts. Given that the scope of negation and distributivity operators blurs the contrast between singular and plural interpretations, we predict that bare nouns can appear in these contexts without having to respect the typicality restriction. It is for this reason that we added NEG/DIS as a variable in our annotation scheme.

Analysis

To track the explanatory value of the variables we annotated for, we need a statistical model that is able to deal with small datasets and multiple interactions. We selected Conditional Inference Trees (CIT, Tagliamonte and Baayen, 2012) as our model. The output of CIT organizes dimensions of variation as a decision tree. What the model does is to check whether the independent variables we select—variables b) through f)—are significantly associated with the response variable in the dataset—variable a), the choice for a bare noun or *yi+N*. If so, it evaluates which of them has the strongest association and uses the outcome of this evaluation to introduce a binary split in the dataset based on the values of the independent variable. These steps are repeated until no further significant associations are found ($\alpha = 0.05$).

Results

Figure 5.1 presents the CIT output we obtained. Two factors that we annotated for turn out to have a significant effect on the distribution of bare nouns and *yi +N*. The first one is typicality (node 1), the second one NEG/DIS (node 2). The data show that there are two contexts in which bare nouns are likely to be used: those in which the verb-noun combination respects the typicality restriction (node 5) and those in which they occur in the scope of negation or a distributivity operator (node 3). In all other contexts, *yi +N* is virtually the only option (node 4). For reference, we note that there are two cases of bare nouns that are part of node 4.

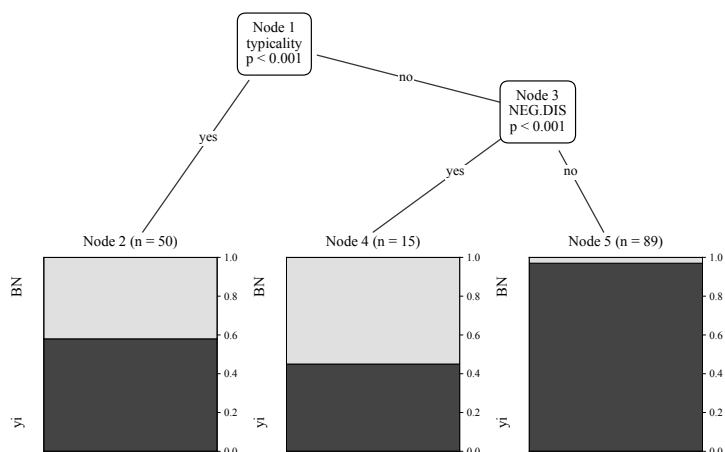


Figure 5.1: Conditional Inference Tree output for the distribution of bare nouns and *yi* + N in all chapters of *Harry Potter and the Philosopher's Stone*

Discussion

The question we addressed with our corpus study was whether bare nouns in indefinite singular contexts are restricted to those in which they stand in a typical relation to their verb. The predictions we make if bare nouns are indeed subject to the typicality restriction are (i) that bare nouns only appear in indefinite singular contexts that respect the typicality restriction and (ii) that *yi* + N is the only option in all other indefinite singular contexts. We argue that these predictions are borne out. Under the assumption that datapoints that get a yes value for NEG/DIS do not qualify as indefinite singular contexts, there are only two exceptions to predictions (i) and (ii), *viz.* the two cases of bare nouns that are part of node 4:

- (21) tā zài sōuxún fēixíng mùbiāo fāngmiàn yǒu zhe **guòrén de jìqiǎo**
 he at search flying target domain have ASP exceptional DE skill
 'He has **an exceptional skill** in searching for flying targets'
- (22) wǒ yǒu **chōngfèn de lǐyóu** shǒu kǒu rú píng
 I have adequate DE reason keep silent (lit. keep mouth as bottle)
 'I have **an adequate reason** to keep silent'

Even though these cases require an explanation, we consider their exceptional status sufficient reason to put them aside for now and move ahead with the clear generalization that bare nouns in Mandarin are sensitive to the typicality restriction.

Three further results of our corpus study deserve to be highlighted. The first is that NEG/DIS has a higher impact on the distribution of bare nouns and *yi* + N than referentiality. This is an important finding, in particular in view of the fact that there are more non-referential contexts than NEG/DIS contexts: if referentiality itself were relevant, we would expect the numbers to be in favor of revealing a bigger role for referentiality. The fact that NEG/DIS has a higher impact then strongly suggests that we are correct in assuming that it targets the distinction between singular and plural indefinite contexts that we designed it for and is independent of considerations about referentiality. The two other results that deserve to be followed up on are the lack of impact of modification and aspect. Given the size of our corpus, we decided not to try and distinguish between different types of modifiers and restricted ourselves to the role of a single aspect marker. We leave it for future research to check whether making further distinctions might reveal new insights about the restrictions on Mandarin bare nouns in indefinite singular contexts.

Conclusion

On the basis of the corpus study we reported on in this section, we conclude that bare nouns in indefinite singular contexts are subject to the typicality restriction: they typically only appear with verbs that either refer to a typical use of the object denoted by the noun or refer to the typical way this object comes into existence. This is an important result as it shows that Mandarin bare nouns in indefinite singular contexts are restricted in a systematic way that corresponds to the intuition of Huang (2015) about the typicality restriction for a smaller set of verb-noun combinations that he considers to involve pseudo-incorporation. We consequently conclude that Mandarin bare nouns in indefinite singular contexts are necessarily pseudo-incorporated. Given that the typicality restriction is similar to the one Dayal (2003, 2011) report for Hindi verb-noun combinations, we also conclude that Mandarin qualifies as a type (iii) language in the typology of pseudo-incorporation we argued for in Section 5.4.

This section concludes the second challenge we set for ourselves, *viz.* to identify a property that all bare nouns in indefinite singular contexts share and that links them to pseudo-incorporation. In Section 5.6, we turn to our third and final challenge, *viz.* to build an analysis that captures the typicality restriction, allowing us to account for Mandarin pseudo-incorporation, while at the same time providing us with the analytical strength to position Mandarin in a broader typology of pseudo-incorporation.

5.6 Analyzing pseudo-incorporation: capturing typicality and the place of Mandarin pseudo-incorporation

In Section 5.5, we argued that Mandarin bare object nouns in singular indefinite contexts are pseudo-incorporated on the basis of the fact that they only appear in verb-noun combinations in which the verb describes a typical way to use the object denoted by the noun or a typical way to bring the object denoted by the noun into existence. We refer to this restriction as a restriction of typicality (see our discussion of Huang in Section 5.2), noting that the same type of restriction is sometimes referred to with different terms (e.g., institutionalization, well-establishedness or proto-typicality). We now move to the third challenge we set for ourselves in Section 5.2, viz. to develop an analysis that derives the typicality restriction and allows us to position Mandarin pseudo-incorporation in a broader typology. Our primary focus will be on the typology we proposed in Section 5.4, viz. one that opposes languages in which there are no clear lexical restrictions on pseudo-incorporation (e.g., Hungarian) to those for which the literature has identified restrictions on the types of verbs (e.g., Spanish) and those in which there are restrictions on verb-noun combinations. Next to Mandarin, the latter type also includes Hindi.

The structure of this section is as follows. We start in 5.6.1 by evaluating Luo's analysis, which is, to our knowledge, the only formal semantic analysis to date that has been worked out in full. The conclusion we will arrive at is that Luo does not derive the typicality restriction and *a fortiori* misses the analytical power to situate Mandarin pseudo-incorporation in a cross-linguistic typology. This will lead us to Dayal's analysis of Hindi pseudo-incorporation in 5.6.2. We will argue that Dayal derives the typicality restriction and that her account for Hindi could in principle be transferred to Mandarin. However, we will also argue that Dayal's account does not come with the analytical power we need to situate Mandarin pseudo-incorporation in a broader typology. With no adequate account available for Mandarin nor for the language that it resembles most qua pseudo-incorporation, it is clear that we cannot rely on existing analyses and have to build up our own. To do so, we first look into Espinal & McNally's analysis of pseudo-incorporation in Spanish and Catalan in 5.6.3. Even though their analysis is not designed to account for restrictions on verb-noun combinations, it contributes a distinction between verbs and verb classes that we will build into our analysis, which we work out in 5.6.4. We conclude this section in 5.6.5 with a discussion of how the analysis we propose complies with the requirements we put forth and

what its implications are for the analysis of pseudo-incorporation in Mandarin and the literature on pseudo-incorporation in general.

As a final point before diving into 5.6.1, we think it is good to establish a basic (theme) argument suppression semantics for pseudo-incorporation verbs that we can use as a basis of comparison for the different proposals we will review. We present this semantics in two notations, a neo-davidsonian (23) and a more traditional one (24). For comparison, we also present what regular verbs look like in these two notations (23a/24a):

(23) regular vs. pseudo-incorporation verbs in neo-davidsonian notation

- a. $\lambda x \lambda y \lambda e [V(e) \& Ag(e) = y \& Th(e) = x]$
- b. $\lambda P_{\langle e, t \rangle} \lambda y \lambda e [V(e) \& Ag(e) = y \& \exists x [P(x) \& Th(e) = x]]$

(24) regular vs. pseudo-incorporation verbs in traditional notation

- a. $\lambda x \lambda y [V(x)(y)]$
- b. $\lambda P_{\langle e, t \rangle} \lambda y \exists x [V(x)(y) \& P(x)]$

In both (23) and (24), the x variable that is abstracted over in the a versions has been existentially closed off in the b versions, formalizing argument suppression. Instead, we find that there is a property P that is abstracted over, where P is predicated of x . For concreteness, we will sometimes work with the running example in (25) that allows for pseudo-incorporation in the different languages we will be discussing:

- (25) Yuēhàn chuān xīzhuāng.
 John wear suit
 ‘John wears **a suit**.’

The $\langle e, t \rangle$ -type semantics of *xīzhuāng* (‘suit’) straightforwardly combines with pseudo-incorporation versions of *chuān* (‘wear’) and after combining the result with the e -type semantics of *Yuēhàn* (‘John’) and other standard operations, the end results look as specified in (26) and (27).

(26) neo-davidsonian notation:

$$\exists e [WEAR(e) \& Ag(e) = y \& \exists x [SUIT(x) \& Th(e) = x]]$$

(27) traditional notation:

$$\exists x [WEAR(x)(y) \& SUIT(x)]$$

Both notations give rise to similar truth conditions: there has to be a suit that is worn by John. The only difference lies in the way the relation between John and the

suit is established. In neo-davidsonian notation, this is achieved indirectly through event participation whereas in traditional notation, the relation is established directly through the semantics of the verb. With the basics in place, we can turn to Luo's analysis of Mandarin pseudo-incorporation.

5.6.1 Luo (2022)

Luo (2022) credits Dayal (2011, 2015) and Schwarz (2014) as the basis for his account. Luo assumes that verbs allowing for pseudo-incorporation come in two guises, a regular one and a pseudo-incorporation one. He proposes the following semantic format for a pseudo-incorporation entry:

$$(28) \quad \lambda x_k \lambda s.t.*\{e|V(e)\&\exists x[Ux_k(x)\&Theme(e) = x]\&e \leq s\}$$

We discuss the entry by comparing it to the one we proposed in (23b). The first—minor—difference between (23b) and (28) is that Luo assumes the agent is not encoded in the verb itself but is added at a later stage through event identification (Kratzer, 1996). The second difference is that he takes incorporation verbs to select kinds (λx_k) instead of properties ($\lambda P_{\langle e,t \rangle}$), a choice motivated by the fact that Luo works in the KA. The third and final difference is that the verb entry in (28), after combining with an object noun, does not give rise to predicates of events but to event kinds, i.e., functions from situations to the largest plurality of events of which V holds and that have an instantiation of the kind corresponding to the noun as theme. Luo argues that this change is relevant to account for the contrast between (29) and (30):

(29) wǒ xǐhuān **diào yú**.
I like catch fish
'I like **fishing**.'

(30) ??wǒ xǐhuān **diào liǎng tiáo yú**.
I like catch two CL fish
Lit.: 'I like catching **two fish**.'

According to Luo, the difference in felicity between (29) and (30) reveals that *diào* can combine with bare nouns to give rise to event kinds but that a similar event kind reading cannot be obtained for *diào* combining with nouns preceded by a numeral and a classifier.

For our present purposes, the main point we want to make about Luo's analysis is that it does not provide an account of the restricted nature of pseudo-incorporation in

Mandarin. Even though Luo indicates that he assumes there could be some flavor of typicality related to the creation of event kinds, the semantic tools he puts to use (pluralization (*), the iota-shift (ι), and abstraction over situations (λs)) do not implement this, independently of whether we maintain that incorporation verbs select predicates or kinds. We conclude that Luo’s analysis does not derive the typicality restriction and that it consequently also does not have the analytical power to situate Mandarin pseudo-incorporation in a cross-linguistic typology. In 5.6.2, we turn to Dayal’s analysis of Hindi and ask whether it fares better at the criteria we set forth for our analysis of Mandarin pseudo-incorporation. With Hindi, we stay within the languages that have been argued to have a typicality restriction on verb-noun combinations.

5.6.2 Dayal (2003, 2011)

There are two versions of Dayal’s analysis of Hindi pseudo-incorporation: a manuscript from 2003 and the updated and extended version published in 2011. In both, Dayal analyzes Hindi pseudo-incorporation as a case of theme argument suppression: whereas regular transitive verbs are given the event-based semantics we saw in (23a), verbs involved in pseudo-incorporation receive the entry in (31):

$$(31) \quad \lambda P_{\langle e,t \rangle} \lambda y \lambda e [P-V(e) \& Ag(e) = y]$$

(31) is different from the more basic entry in (23b) in that P is not predicated of an x that is existentially closed off but is added in front of V . In Dayal (2003), the condition $P-V(e)$ remains undefined but a condition requiring that the event e be appropriately classificatory is added (32). This condition is central to Dayal’s account of the typicality restriction in Hindi.

$$(32) \quad \lambda P_{\langle e,t \rangle} \lambda y \lambda e [P-V(e) \& Ag(e) = y \& \textit{appropriately - classificatory}(e)]$$

Where an event denoted by a predicate δ that incorporates a property γ is appropriately classificatory iff

$$\diamond \textit{probable}(\exists e[\delta(e) \& \exists y[Ag(e) = y] \& \exists x[\gamma(x) \& Th(e) = x]])$$

In the published version from 2011, Dayal does define $P-V$ and leaves out the *appropriately classificatory* condition:

$$(33) \quad \lambda P_{\langle e,t \rangle} \lambda y \lambda e [P-V(e) \& Ag(e) = y]$$

Where $\exists e[P-V(e)] = 1$

$$\textit{iff} \exists e'[\textit{CATCH}(e') \& \exists x[P(x) \& Theme(e') = x]]$$

Given that we are interested in the notion *appropriately classificatory* in (32) but at the same time recognize the need to define *P-V* as in (33), we discuss Dayal's analysis as a merger of (32) and (33):

- (34) $\lambda P_{\langle e,t \rangle} \lambda y \lambda e [P-V(e) \& Ag(e) = y \& \textit{appropriately-classificatory}(e)]$
 Where $\exists e [P-V(e)] = 1$ iff
 $\exists e' [catch(e') \& \exists x [P(x) \& Theme(e') = x]]$
 AND an event denoted by a predicate δ that incorporates a property γ is appropriately classificatory iff
 $\diamond probable(\exists e [\delta(e) \& \exists y [Ag(e) = y] \& \exists x [\gamma(x) \& Th(e) = x]])$

By adding the *appropriately classificatory* condition at the level of the event and spelling out the truth conditions of $\exists e [P-V(e)]$ as involving both the verb and the noun, Dayal derives the fact that it is the combination of the verb and the noun that matters in the restrictions on pseudo-incorporation in Hindi. As such, she has an account of the typicality restriction that we could, *a priori*, adopt in our analysis of Mandarin pseudo-incorporation. However, the *appropriately classificatory* condition is problematic. First, the exact formulation is too weak, any utterance including a non-negated version of (12) automatically meeting the condition. Indeed, under the assumption that speakers observe Grice's maxim of quality, an utterance stating that there is an event of which *laRkii-dekhnaa* ('girl-see') holds will automatically lead the hearer to consider it probable for there to be an event of which *laRkii-dekhnaa* holds, rendering the restriction of the *appropriately classificatory* condition obsolete. One could of course work on the exact formulation (see also Dayal, 2011, 2015), but this would not solve the second and more serious problem, viz. the fact that the *appropriately classificatory* condition in (34) lacks linguistic grounding. The role of linguistic grounding can best be captured in two questions about Hindi pseudo-incorporation that Dayal's analysis does not provide an answer to and that would carry over to Mandarin pseudo-incorporation:

- (i) Why is it that every pseudo-incorporation verb in Hindi comes with the same *appropriately classificatory* condition?
- (ii) Why is it that pseudo-incorporation verbs in some other languages do not come with this condition?

The crux of linguistic grounding is this: for an analysis to reach explanatory adequacy, it needs to derive which verbs are sensitive to the *appropriately classificatory*

condition and which are not. Merely stating that every pseudo-incorporation verb in Hindi comes with the appropriately classificatory condition is a statement of fact, not the derivation of this fact. And in the light of the typological perspective we presented in Section 5.4, it is equally important to explain why the appropriately classificatory condition appears in the entry of pseudo-incorporation verbs in Hindi and Mandarin but not on those in Hungarian. We conclude that Dayal’s appropriately classificatory condition can account for the typicality restriction we find in Hindi and Mandarin but that Dayal’s analysis does not provide us with the analytical power we need to position Hindi—and *a fortiori* Mandarin—in a cross-linguistic typology of pseudo-incorporation. The more general conclusion after our discussion of Luo (2022) and Dayal (2003, 2011) is that there is no satisfactory account of pseudo-incorporation in languages with a central role for a typicality restriction. We consequently cannot but build up our own analysis. Before doing so in 5.6.4, we turn to the analysis Espinal and McNally (2011) propose for pseudo-incorporation in Spanish and Catalan. Even though the languages they work on are *prima facie* different from Mandarin, they provide a specific view on how to formalize one aspect of linguistic grounding that we will be relying on in our analysis.

5.6.3 Espinal & McNally (2011)

We discuss the account proposed by Espinal and McNally (2011) and take (10), repeated here in its Spanish version, as our running example:

- (35) Juan **llevaba traje**.
 John wore suit
 ‘The man wore **a suit**.’

According to Espinal and McNally (2011), *llevar* is part of a cluster of verbs known as HAVE-verbs that we can apply the following lexical rule to:²

- (36) **Input:** $\lambda y \lambda e [V(e) \& \theta(e) = y \& \exists e' [depend(e, e') \& have(e') \& havee(e') = y]]$
Output: $\lambda e [V(e) \& \exists e' [depend(e, e') \& have(e') \& havee(e') = \theta(e)]]$

The effect of this lexical rule can be seen from a comparison of the bolded parts: whereas the input contains abstraction over the variable *y* that is identified as the theme

²For expository reasons, we do not reproduce the intensional elements of the rule. Our discussion is independent of these elements.

argument of e (compare with (23a)), this abstraction has disappeared from the output. Espinal & McNally’s analysis thus conforms to the one in (23b) in that it takes pseudo-incorporation to involve theme argument suppression. The non-bolded part of the input and output is identical and is intended to capture the fact that the rule is limited to HAVE-verbs. *Llevar* (‘wear’) qualifies in the sense that for someone to wear something like a suit, the suit should—at least temporarily—be in that person’s possession.

As the reader will have noticed, the argument structure of the output in (36) differs from the one of pseudo-incorporation verbs in (23b) in that there is no way to directly combine the noun with the verb through function application. Espinal & McNally assume the combination of the two is based on the intersective composition rule defined in (37):

- (37) If $\llbracket V \rrbracket = \lambda e[V(e)]$ and θ is an implicit role function defined for V ,
 and if $\llbracket N \rrbracket = N$, a property
 then $\llbracket [vV N] \rrbracket = \lambda e[V(e) \& N(\theta(e))]$

(37) relies on the intuition that theme argument suppression does not change the way the events V applies to are conceived: *llevar* is still conceived as an event that brings together a *wearer* and a *wearee*, and the output of *llevar* after applying the lexical rule in (36) thus conforms to the condition in the first line of (37). Combining *llevar* with *traje* (‘suit’) and composing the result with a standard semantics of *Juan* through event identification leads to the final semantics of (35):

- (38) $\exists e[\text{WEAR}(e) \& \text{Ag}(e) = \text{J} \& \text{SUIT}(\theta(e))]$

(38) states that Juan is the agent of an event of wearing, whose theme argument has the property of being a suit.

The ingredient of Espinal & McNally’s analysis of Spanish/Catalan pseudo-incorporation that we will transfer to our analysis is the general format of a lexical rule. Lexical rules were introduced into semantics to account for semi-productive generalizations within the lexicon, defining the possible range of a generalization without fixing its actual range (Dowty, 1979). Espinal & McNally use lexical rules to restrict the possible range of pseudo-incorporation to those verbs that have a HAVE-component without fixing that every verb of the relevant type can always function as a pseudo-incorporation verb. In our account, we will use lexical rules to implement the typicality restriction.

We note that lexical rules provide a level at which we can derive linguistic grounding but that they do not guarantee it. In this respect, we acknowledge that the lexical

rule of Espinal & McNally formally derives that individual HAVE-verbs can function as pseudo-incorporation verbs in Spanish and Catalan. However, the rule has nothing to say about why there are no parallel rules for other verb classes. As such, it merely stipulates that the class of HAVE-verbs has an exceptional position and does not actually provide linguistic grounding. We get back to this point in 5.6.5 when we discuss the implications of our analysis for the broader semantic literature on pseudo-incorporation.

5.6.4 Our analysis

The analysis we will develop here builds on the one that Le Bruyn et al. (2016) proposed for pseudo-incorporation in languages like Spanish and Catalan. For expository purposes, we first work out our analysis for Mandarin pseudo-incorporation and get back to the main differences with Le Bruyn et al. at the end. The presentation of the analysis itself is organized in three stages. In the first, we lay the groundwork we need and propose a preliminary analysis focused on individual verb entries. In the second stage, we lift the analysis to the level of a lexical rule and in the third and final stage, we sketch a number of *caveats* and possible extensions. Our focus throughout this section is on capturing the typicality restriction. In 5.6.5 we will argue that our proposal also provides us with the analytical power needed to define the position of pseudo-incorporation in Mandarin in a broader typology of pseudo-incorporation.

The groundwork

We propose (39) as a model entry for pseudo-incorporation verbs in Mandarin, adopting the traditional notation from (24b):

$$(39) \quad \lambda P_{\langle e, \langle e, t \rangle \rangle} \lambda y \exists x [V(x)(y) \& P(y)(x)]$$

(39) formalizes pseudo-incorporation as an instance of (object) argument suppression, in line with the analyses we reviewed in 5.6.1 to 5.6.3. What sets (39) apart specifically from Dayal's analysis is that it selects a relational (type $\langle e, \langle e, t \rangle \rangle$) instead of a sortal predicate (type $\langle e, t \rangle$) and that it does not come with a restriction along the lines of the appropriately classificatory condition. In what follows, it will become clear that these differences are related, the type requirement allowing us to derive the typicality restriction that the appropriately classificatory condition is meant to capture.

To appreciate the effect of (39), we use it to work out the running example we proposed in (25), repeated below as (40):

- (40) Yuēhàn chuān xīzhuāng.
 John wear suit
 ‘John wears **a suit**.’

For the semantics of *chuān* (‘wear’), (40) leads to the following pseudo-incorporation entry:

$$(41) \lambda P_{\langle e, \langle e, t \rangle \rangle} \lambda y \exists x [\text{WEAR}(x)(y) \& P(y)(x)]$$

For the semantics of *xīzhuāng* (‘suit’), we assume that it is a sortal noun that can be turned relational by accessing its qualia structure (see also Vikner and Jensen, 2002). The idea behind QUALIA structure is that the lexicon provides information about nouns that extends beyond classical entries, conforming to a general format consisting of four perspectives on objects, known as roles (Pustejovsky, 1995). One of these roles is the *telic* role, specifying – where applicable – what the objects denoted by the noun are designed for. For *xīzhuāng*, we assume the telic role is specified as *y wears x*, leading us to the relationalized interpretation in (42):

$$(42) \lambda y \lambda x (\text{SUIT}(x) \& \text{WEAR}(x)(y))$$

With (41) and (42) in place, the semantics of *chuān xīzhuāng* is straightforwardly derived as in (43). Combining it with the semantics of *John*, we obtain the final semantics of (40) as specified in (44).

$$(43) \lambda P_{\langle e, \langle e, t \rangle \rangle} \lambda y \exists x [\text{WEAR}(x)(y) \& P(y)(x)] \lambda y \lambda x (\text{SUIT}(x) \& \text{WEAR}(x)(y)) \\ \lambda y \exists x [\text{WEAR}(x)(y) \& \text{SUIT}(x) \& \text{WEAR}(x)(y)]$$

$$(44) \exists x [\text{WEAR}(x)(y) \& \text{SUIT}(x) \& \text{WEAR}(x)(y)]$$

(44) states that there is a wearing relation that holds between Yuehan and a suit. These are the desired truth conditions. We note that the relational information is specified twice, once stemming from the semantics of the verb, and once stemming from the relational semantics of the noun.

Rather than considering the doubling of the relational information in (44) a disadvantage, we consider it a crucial asset of the analysis we have built up so far. We argue why on the basis of the overview of the four QUALIA roles in (45). The overview provides typical characterizations of the different roles as well as examples accompanied by their predicate logical representation.

- (45) Overview of the four qualia roles

The formal role specifies the position of a noun within a taxonomy. A noun like *book* could, e.g., be classified as an artifact.

artifact (x)

The constitutive role specifies what the objects denoted by the noun consist of. A book can, e.g., be said to consist of, e.g., a cover.

x consists_of y, with cover (y)

The telic role specifies what the objects denoted by the noun are designed for. A book can, e.g., be said to be designed to be read by someone.

y reads x

The agentive role specifies the creator denoted by the noun. A book can, e.g., be said to have been written by someone.

y wrote x

We make two observations on the basis of the overview in (45). The first is that only three roles provide relational information in which the object denoted by the noun is said to stand in a relation to other objects or individuals: the *constitutive*, *telic* and *agentive* roles. The second observation is that of the three remaining roles, only two provide relations in which the individual that the object denoted by the noun is said to stand in a relation to, appears in the subject position. The upshot of this round of elimination is that only two qualia roles can be used to relationalize sortal nouns in such a way that they double the relational information of verbs that they appear as objects of: the *telic* and the *agentive* roles. Going back to what typical verb-noun combinations are, we argue that these can be analyzed as involving exactly those relations we find in the *telic* and the *agentive* roles of nouns, the former describing the typical way the objects are used, the latter the typical way the objects have come into existence. The conclusion that imposes itself is that a semantics that can capture the restriction of pseudo-incorporation to verb-noun combinations in which the verb doubles the relational information in the noun's *telic* or *agentive* roles, succeeds in directly capturing the typicality restriction. In what follows, we argue that we can accomplish by upgrading our model entry in (39) to a lexical rule.

From a model entry to a lexical rule

Based on the model entry in (39) and the discussion above, we propose the lexical rule in (46):

- (46) **Input:** $\lambda x \lambda y (V(x)(y))$
Output: $\lambda P_{\langle e, t \rangle} \lambda y \exists x [(REL_{qv}(P))(y)(x)]$

Where REL_{qv} is an operation that transforms sortal nouns into relational ones based on their qualia structure iff the relational information there matches that of the verb in the input. REL_{qv} is undefined in the absence of a match.

The lexical rule in (46) can in principle apply to any transitive verb but will only yield a relevant output for those verbs whose semantics is doubled by the relational information included in the QUALIA structure of a set of nouns. There are two differences between the output in (46) and the model entry in (39). The first is that P is no longer of type $\langle e, \langle e, t \rangle \rangle$, the relational interpretation of P being created in the course of the derivation through REL. This does justice to the fact that a noun like *suit* is considered a sortal noun. The second difference is that the verb is no longer directly encoded in the output of (46). The semantics of the verb is indirectly encoded, though, by making the relationalization operation dependent on a match between the relational information that obtains and the relational information of the verb. This not only ensures that there is no more doubling of relational information but also captures the typicality restriction we are after: pseudo-incorporation that is based on verbs derived by the lexical rule in (46) only allows for verb-noun combinations in which the verb describes the typical way the objects denoted by the noun are used or the typical way these objects have come into existence.

For concreteness, we work out our running example, starting from the regular semantics of *chuān* and *xīzhuāng* in (47) and (48):

- (47) $\lambda x \lambda y \text{WEAR}(y)(x)$
(48) $\lambda x \text{SUIT}(x)$

Applying the lexical rule in (46) to (47), we obtain the following entry for *chuān*:

- (49) $\lambda P_{\langle e, t \rangle} \lambda y \exists x [(REL_{\text{WEAR}}(P))(y)(x)]$
Where REL_{WEAR} transforms a sortal predicate P into $\lambda y \lambda x (P(x) \& \text{WEAR}(x)(y))$ if the wear relation is encoded in the telic role of the noun corresponding to P. REL_{WEAR} is undefined in the absence of a match.

Under our assumption that the telic role of *xīzhuāng* is specified as *y wears x*, (49) can combine with (48). The full derivation is worked out in (50):

- (50) $\lambda P_{\langle e, t \rangle} \lambda y \exists x [(REL_{\text{WEAR}}(P))(y)(x)] \quad \lambda x \text{SUIT}(x)$
 $\lambda y \exists x [(REL_{\text{WEAR}}(\lambda x \text{SUIT}(x)))(y)(x)]$ (function application)

$$\begin{array}{ll}
 \lambda y \exists x [\lambda y \lambda x (\text{SUIT}(x) \& \text{WEAR}(x)(y))(y)(x)] & \text{(REL)} \\
 \lambda y \exists x [\text{SUIT}(x) \& \text{WEAR}(x)(y)] & y \quad \text{(function application)} \\
 \exists x [\text{SUIT}(x) \& \text{WEAR}(x)(y)] & \text{(function application)}
 \end{array}$$

The final line in (50) shows how the lexical rule in (46), combined with our earlier assumptions about the qualia structure of *xīzhuāng* gives rise to the same truth conditions as we found in (44) on the basis of our model entry in (39).

Caveats and extensions

With (46), we have formalized what we consider to be the core semantics of pseudo-incorporation for languages that have been argued to have lexical restrictions on verb-noun combinations. We however want to make two further observations. The first is merely to remind the reader of the fact that lexical rules were designed to capture generalization in the lexicon that are not necessarily fully productive. We consequently take (46) to formalize the prototypical range of pseudo-incorporation in Mandarin but not its actual range. The second observation is that we think slight variations on the rule in (46) are to be expected, subject to intra- and cross-linguistic variation. The first is that the relational semantics of verbs might match the one in the telic or agentive roles of nouns while still being more specific. E.g., the noun *shū* ('book') in Mandarin can combine with *kàn* and *dú*, both resulting in a meaning along the lines of reading a book. We do not want to exclude that *kàn shū* and *dú shū* still have slightly different meanings, though, and we consequently assume that the semantic import of verbs could go beyond the matching function they have in the output of (46). The second extension consists in relaxing the constraint in (46) according to which only relations included in qualia structure play a role. This seems reasonable for those relations that have independently been argued to play a role in language, possession in particular (see Vikner and Jensen, 2002). The third and final extension we foresee is one in which verbs can show a similar interaction with nouns. A case in point could be the way a semantically empty verb like relational have in the sense of Partee (1999) combines with a relational noun like *child* to derive that the subject is a parent of the child. The underlying composition is similar to the one we pursue in (46), *viz.* one in which the subject ends up as an argument of the relation encoded in the object noun.

A comparison with Le Bruyn et al. (2016)

The main goal of this section was to build up an analysis of Mandarin pseudo-incorporation that captures the typicality restriction. We have argued that we have achieved

this result with the lexical rule in (46). The theoretical intuition underlying (46) is that there is a link between the relational information in the qualia structure of nouns and the possible range of pseudo-incorporation. This intuition is due to Le Bruyn et al. (2016) and we could consequently have decided to present our analysis as a variant of Le Bruyn et al.'s. For expository reasons we have opted not to do so: the analysis of Le Bruyn et al. focuses on languages like Spanish and is embedded in a larger discussion on relationality, possession and dynamic semantics that would distract too much from the central issue here, *viz.* the derivation of the typicality restriction. We do want to highlight two differences between the analyses that are relevant for our discussion of the place of Mandarin pseudo-incorporation in 5.6.5.

Formally speaking, the most important difference is that we have simplified the analysis of Le Bruyn et al. by relegating the relationalization of nouns to the realm of pragmatics instead of keeping it in the semantics. For the interested reader, we note that the simplification consists in relying on a relationalization operator rather than on an explicitation operator, the latter requiring a move to dynamic semantics.

The formal difference is, in turn, inspired by a conceptual one. Le Bruyn et al.'s analysis is focused on arguing in favor of a unified analysis of relational HAVE and pseudo-incorporation with HAVE-verbs. One of the goals of their analysis is consequently to highlight the similarity between relational and sortal nouns. With our focus on Mandarin and the typicality condition, the role of relational nouns and HAVE-verbs was not a central concern. What is interesting, though, is that the same theoretical intuition can cover a range of phenomena. In 5.6.5, we reflect on this in the light of the typology of pseudo-incorporation we argued for in Section 5.4.

5.6.5 Discussion

In this section, we tackled the third and final challenge we set for ourselves, *viz.* to develop an analysis of Mandarin pseudo-incorporation that derives the typicality restriction and has the analytical power to account for the place of Mandarin pseudo-incorporation in a cross-linguistic perspective. In 5.6.4 we already argued that our analysis derives the typicality restriction. Here, we focus on the question whether it also has the analytical power needed to account for the place of Mandarin pseudo-incorporation in the face of cross-linguistic variation. We approach this question at two levels. The first is the level of the typology we introduced in Section 5.4, distinguishing between (i) languages like Hungarian where pseudo-incorporation is virtually unrestricted, (ii) languages like Spanish with a restriction at the level of verbs, and

(iii) languages like Mandarin and Hindi that share the typicality restriction on verb-noun combinations. The second level is that of the semantic properties we discussed in Section 5.3.

Restrictions on pseudo-incorporation

In 5.6.2, we argued that Dayal (2003, 2011) was able to derive the typicality restriction but crucially lacked the analytical power to position Hindi in a broader typology of restrictions on pseudo-incorporation. Here, we argue that our analysis fares better.

Differently from Dayal, the typicality restriction in our analysis does not depend on the addition of the meta-linguistic *appropriately classificatory* condition but on the actual way verbs and nouns interact. As such, our account of the typicality restriction is linguistically grounded, allowing us to explain why it applies to all verbs that participate in pseudo-incorporation in Mandarin, but also putting us in a better position to shed light on cross-linguistic variation.

To account for the lack of restrictions on verb-noun combinations in Hungarian, we postulate that Hungarian pseudo-incorporation involves a special mode of composition – e.g., as formalized in Farkas and Swart (2003) – but that it does not involve the same type of verb-noun interaction as the one we find in Mandarin. On this analysis, there is no *a priori* reason to expect Hungarian pseudo-incorporation to be subject to the same type of typicality restriction as Mandarin pseudo-incorporation.

For the type of pseudo-incorporation we find in languages like Spanish, we think our analysis suggests that there is more overlap with the type of pseudo-incorporation in Mandarin than the typology we presented in Section 5.2 suggests. The claim in the literature is that there is a clear restriction to HAVE-verbs in Spanish but that there is no restriction on object nouns. This seems – at first sight – to be a different situation than the one we find in Mandarin. Suppose, however, that Spanish pseudo-incorporation were to be analyzed with the variant of our analysis in which the matching condition in (46) is relaxed and allows for – or even specializes in – possessive relations. In that case, the lack of restrictions on nouns could very well be argued to be a perceived rather than an actual one, owing to the fact that possession can be inferred for most any object (Vikner and Jensen, 2002). The real difference between Mandarin and Spanish pseudo-incorporation would then be the specific type of relations the two specialize in. Zooming out further, pseudo-incorporation as we have formalized it in (46) could allow for slightly different restrictions across languages, depending on whether it specializes in the relations included in QUALIA structure, in possessive

relations, in the relations included in relational nouns, or potentially language-specific mixes of the preceding. This suggests that the three-way typology we presented in Section 5.2 might well underlyingly be a two-way typology, opposing languages with virtually no restrictions on pseudo-incorporation like Hungarian to those with restrictions like Mandarin, Hindi and Spanish-like languages. As suggested above, pseudo-incorporation in the former would be based on a special mode of composition whereas in the latter, pseudo-incorporation would follow the analysis we proposed in (46).

At this point, it is instructive to go back to the two exceptions we found in our corpus study in Section 5.5 and interpret them in the light of our suggestion that the types of pseudo-incorporation we find in Spanish and Mandarin might be closer to one another than the typology proposed in 5.4 suggested. For convenience, we repeat these exceptions here:

- (51) tā zài sōuxún fēixíng mùbiāo fāngmiàn yǒu zhe **guòrén de jìqiǎo**
 he at search flying target domain have ASP exceptional DE skill
 ‘He has **an exceptional skill** in searching for flying targets’
- (52) wǒ yǒu **chōngfèn de lǐyóu** shǒu kǒu rú píng
 I have adequate DE reason keep silent (lit. keep mouth as bottle)
 ‘I have **an adequate reason** to keep silent’

The exceptional status of (51) and (52) resided in the fact that *yǒu* and *jìqiǎo* and *yǒu* and *lǐyóu* do not stand in a typicality relation. The typicality criterion would consequently require the addition of numeral-*yi* for *jìqiǎo* and *lǐyóu*, and this is not what we find. However, in the light of our suggestion that the types of pseudo-incorporation we find in Spanish and Mandarin are not fundamentally different, the fact that (51) and (52) involve the verb *yǒu* (‘have’) does not need to come as a surprise. Indeed, what (51) and (52) suggest is that pseudo-incorporation in Mandarin is not only limited to relations in qualia structure, but also allows for possessive relations, at least for certain verb-noun combinations.

Semantic properties of pseudo-incorporated nouns

In 5.3, we made an inventory of the semantic properties that are typically associated with pseudo-incorporated nouns: (i) narrow scope, (ii) reduced discourse transparency, (iii) restricted modification possibilities, and (iv) number neutrality. We also argued that only the first has been argued to be stable across languages whereas properties (ii) to (iv) tend to vary. As the reader can easily verify, our analysis derives (i) by ensuring that existential quantification over the object variable is built into the semantic

contribution of the verb. Variants of this strategy are commonplace in the literature on pseudo-incorporation. Our analysis currently does not derive properties (ii) to (iv) and this, we argue, is as it should be in view of the intra-linguistic variation Luo (2022) already hinted at (see Section 5.2). That said, we do want to present two further reflections, moving from Mandarin to the cross-linguistic level.

For Mandarin, we submit that properties (ii) to (iv) may well pop up through pragmatic competition with the numeral-*yi*, and may even become entrenched for specific verb-noun combinations. A case in point is the reduced discourse transparency that – we assume – translates into the low recurrence rate of discourse referents introduced by bare nouns (see Section 5.2). Moving to the cross-linguistic level, we expect the variation we allow for in Mandarin to also present itself in languages in which pseudo-incorporation follows the analysis in (46). Assuming that pseudo-incorporation in Spanish-like languages fits the bill, this is in line with Le Bruyn et al.’s observation that properties (ii) to (iv) are not stable in these languages (see Section 5.3). Given that we assume Hungarian pseudo-incorporation is based on a different mode of composition, pragmatic competitions have to be worked out separately and we have no grounds to make predictions about the stability of properties (ii) to (iv).

5.7 Conclusion

This chapter set out to resolve the indefinite part of the alternation challenge. Our investigation confirmed our central hypothesis: the distribution of bare nouns in indefinite object positions is systematically constrained by a typicality relation with the verb, a key diagnostic for pseudo-incorporation in Mandarin. Our corpus study provided empirical support for this claim, demonstrating that bare nouns appear almost exclusively in verb-object combinations with typicality.

This finding provides a principled division of labor that resolves the empirical and theoretical questions for the indefinite domain posed in Section 1.3. Empirically, the division of labor between numeral-*yi* and bare nouns is systematic: numeral-*yi* functions as the indefinite article in regular argument positions, while bare nouns are restricted to non-argumental, pseudo-incorporation contexts. Theoretically, our formal analysis captures this division of labor within a modified Properties Approach. By treating pseudo-incorporation as a lexical rule on the verb that is sensitive to the noun’s QUALIA structure, we show how bare nouns are licensed via a mechanism distinct from the standard existential type-shift. This preserves the Blocking Principle: numeral-*yi*, as an overt indefinite article, blocks the covert type-shift for bare nouns

in regular argument positions, forcing them into the alternative pseudo-incorporation structure to express indefiniteness.

With the indefinite side of the alternation puzzle now accounted for, our focus shifts to the definite domain. The co-existence of bare nouns and demonstratives, first established in Chapter 2 and validated in Chapters 3 and 4, presents a distinct set of challenges, and it is not clear of any systematic division of labor between bare nouns and demonstratives. To resolve this, we will proceed in two steps. First, in Chapter 6, published as an independent paper, we will test the most prominent hypothesis in the literature for the competition between bare nouns and demonstratives in definite contexts, namely Jenks (2018) proposal of a weak vs. strong definiteness distinction. Chapter 6 will in fact show that Jenks's hypothesis is not supported by empirical data from the translation corpus. Therefore, Chapter 7 will force us to consider a second possibility that there is no genuine semantic alternation, and that the Mandarin demonstrative is in fact functioning as a canonical demonstrative, rather than a marker of definiteness.

CHAPTER 6

Translation Mining: Definiteness across Languages ¹

6.1 Introduction

The distinction between weak and strong definiteness was originally proposed by Schwarz (2009) to account for the difference between German contracted and uncontracted definite articles. Jenks (2018) extends it to Mandarin, linking bare nouns to weak definiteness and demonstratives to strong definiteness.

We present a parallel-corpus study that compares the distribution of German contracted/uncontracted articles and Mandarin bare nouns/demonstratives. The work by Schwarz and Jenks leads us to predict that German contracted articles pattern with Mandarin bare nouns and German uncontracted articles with Mandarin demonstratives. We show that these predictions are only partly borne out and argue for a more fine-grained typology of definiteness and of strong definiteness in particular.

The chapter is structured as follows. After giving relevant background on weak and strong definiteness (Section 6.2), we present our parallel-corpus study (Section 6.3 and Section 6.4) and argue for a new dimension in the typology of definiteness

¹*This chapter was originally published with the title Translation Mining: Definiteness across languages (a reply to Jenks 2018) in *Linguistic Inquiry*, Vol. 53 and is joint work with David Bremmers and Bert Le Bruyn. The original paper addresses the empirical claims made by Jenks (2018). This paper serves the same function in the overall argumentation of the thesis. Please see Section 1.5 for details.*

(Section 6.5). We conclude with a brief summary of the main findings (Section 6.6).

6.2 Weak and strong definites across languages: setting the stage

Here, we give a brief overview of Schwarz (2009) and Jenks (2018), distinguishing between data (Section 6.2.1) and analysis (Section 6.2.2). We conclude with a summary and outlook (Section 6.2.3).

6.2.1 Data

Schwarz (2009) brings together insights from Hawkins (1978)'s seminal work on English definites and from a rich descriptive tradition on definites in other Germanic languages and dialects (e.g., Ebert, 1971a,b). Schwarz's main data are taken from Standard German (see Wiltschko, 2013 on Austro-Bavarian German and Meier, 2019 on Zurich German).

What sets Standard German apart from English is that it has two forms for the definite article: one rendering uniqueness—or “weak” definiteness—and one rendering familiarity—or “strong” definiteness. Weak/Uniqueness definites are primarily the immediate- and larger-situation definites in Hawkins's typology, and strong/familiarity definites predominantly correspond to Hawkins's anaphoric definites. As for the associative (or bridging) uses Hawkins discusses, Schwarz argues that some qualify as strong and others as weak.

The weak vs. strong distinction in Standard German manifests itself formally in that weak definite articles contract with certain prepositions, whereas strong definite articles resist contraction. Outside the prepositional domain, no formal difference can be detected. Key examples from Schwarz (2009) are given in (1) and (2).

- (1) Der Empfang wurde **vom** / **#von dem Bürgermeister** eröffnet.
 the reception was by.the / by the mayor opened
 ‘The reception was opened **by the mayor**.’ (Schwarz, 2009, 40)
- (2) In der New Yorker Bibliothek gibt es ein Buch über Topinambur.
 in the New York library exists there a book about topinambur
 Neulich war ich dort und habe **#im** / **in dem Buch** nach einer Antwort auf
 recently was I there and have in.the / in the book for an answer to
 die Frage gesucht, ob man Topinambur grillen kann.
 the question searched if one topinambur grill can

‘In the New York Public Library, there is a book about topinambur. Recently, I was there and searched **in the book** for an answer to the question of whether one can grill topinambur.’ (Schwarz, 2009, 30)

In (1), the mayor has not been introduced before but is the unique mayor of the contextually salient town. This is a weak/uniqueness context, and the definite article contracts with the preposition. In (2), a book is introduced in the first sentence and referred back to in the second. This is a case of strong/familiarity definiteness, and contraction is not allowed.

Several studies have followed up on Schwarz (2009) and have argued that the weak vs. strong distinction underlies definiteness paradigms in typologically diverse languages (see, e.g., Arkoh and Matthewson, 2013 for an extension to Akan, and Aguilar-Guevara et al., 2019 for an overview). We focus on the case of Mandarin as presented by Jenks (2018).

Jenks provides the following key examples:

- (3) (#Nà / #Zhè ge) **Táiwān (de) zǒngtǒng** hěn shēngqì.
 that / this CL Taiwan (’s) president very angry
 ‘**The president of Taiwan** is very angry.’ (Jenks, 2018, 507)
- (4) Jiàoshì lǐ zuò-zhe yī ge nánshēng hé yī ge nǚshēng. Wǒ zuótiān
 classroom in sit-ASP one CL boy and one CL girl. I yesterday
 yùdào #(nà ge) **nánshēng**.
 meet #(that CL) boy.
 ‘There are a boy and a girl sitting in the classroom. I met **the boy** yesterday.’
 (Jenks, 2018, 510)

(3) shows that nouns referring to unique individuals like the president of Taiwan typically occur bare and that demonstratives are not allowed with them. (4) shows that the bare noun *nánshēng* cannot refer back to the boy in the first sentence and that the demonstrative is required. These facts build a strong case in favor of an active weak/strong distinction for Mandarin bare nouns and demonstratives, parallel to what we find with contracted and uncontracted definite articles in Standard German.

6.2.2 Analysis

The basics of an analysis of weak and strong definites go back to the semantics Schwarz (2009) proposes for each of them.

- (5) The *weak/strong distinction* in Schwarz (2009)

a. Weak definite:

$$\lambda s_r. \lambda P : \exists! x(P(x)(s_r)). \iota x[P(x)(s_r)]$$

b. Strong definite:

$$\lambda s_r. \lambda P. \lambda y : \exists! x(P(x)(s_r) \& x = y). \iota x[P(x)(s_r) \& x = y]$$

Both weak and strong definites are linked to a pragmatically supplied resource situation formalized as the situation pronoun s_r in which the referent is unique. This uniqueness is spelled out in both the presuppositional and the asserted content. Strong definites are special in that they come with a pragmatically supplied index y that the referent of the definite is said to be identical to. This formalizes anaphoricity.

Schwarz assumes the situation pronoun s_r can stand for a contextually salient situation but can also be identified with the (Austinian) topic situation or be bound by a quantifier over situations. In the remainder of this article, we will focus on examples in which s_r is identified with the topic situation—that is, the situation an utterance is about. We follow McKenzie (2012, 2015) in assuming that a situation can consist of multiple eventualities as long as a coherent relation can be established between them. In line with McKenzie’s observations on Kiowa, we take spatiotemporal contiguity to be a good predictor for which eventualities can be considered to belong to the same situation.

Jenks’s analysis of the weak/strong distinction builds on Schwarz’s. We compare Jenks’s entries in (6) with Schwarz’s in (5).

(6) The weak/strong distinction in Jenks (2018)

a. Weak definite:

$$\lambda s_r. \lambda P : \exists! x(P(x)(s_r)). \iota x[P(x)(s_r)]$$

b. Strong definite:

$$\lambda s_r. \lambda P. \lambda Q : \exists! x(P(x)(s_r) \& Q(x)). \iota x[P(x)(s_r)]$$

There are two differences between the entries in (5) and (6). The first is minor and is concerned with the semantic type of the index argument in the strong definite: in Schwarz’s analysis, the index argument y is of type e whereas in Jenks’s analysis, the index argument Q is of type $\langle e, t \rangle$. This move is motivated under the assumption that the index occupies a predicative position in Mandarin (Zhang, 2015).

The second difference is more fundamental. Whereas in (5b) the index argument is active in both the presupposition and the assertion, in (6b) it only appears in the presupposition. The main consequence of this move is that the weak and strong definites become identical in their assertive content while a stronger presupposition is retained

for the strong definite. Under Maximize Presupposition (Heim, 1991), this means that the strong definite has to be used as soon as its presuppositions are met. Jenks (2018, 524) formalizes this in a principle he terms *Index!*.

- (7) *Index!*
Represent and bind all possible indices.

Index! requires the use of strong definites as soon as an anaphoric relation can be established, effectively blocking the use of weak definites in anaphoric contexts. With *Index!* in place, Jenks's analysis is more restrictive than Schwarz's. Whereas Schwarz assumes the difference in acceptability between the contracted and uncontracted forms in (2) is a matter of preference, Jenks hardwires the difference into his analysis.

6.2.3 Summary and outlook

Schwarz (2009) and Jenks (2018) define two types of definiteness and argue that (Standard) German and Mandarin formally distinguish between them. Weak definiteness is concerned with uniqueness and is marked by contracted definites in German and bare nouns in Mandarin. Strong definiteness adds dynamicity by requiring referents to be familiar from previous discourse. Strong definiteness is marked with uncontracted definites in German, whereas Mandarin relies on demonstratives. At the level of analysis, the main difference between Schwarz's and Jenks's analyses lies in the way they deal with the competition between the two types of definiteness. Whereas Schwarz leaves open how the competition should play out, Jenks's constraint *Index!* categorically bans the use of markers of weak definiteness for previously introduced referents.

In Section 6.3 and Section 6.4, we present a parallel-corpus study that puts the predictions of Schwarz's and Jenks's analyses to the test. Parallel—or translation—corpora render the same content in different languages. They are thus particularly suited for checking whether a semantic category that is active in one language is replicated in another. For weak and strong definiteness, the main prediction that follows from Schwarz's and Jenks's work combined is that German contracted definites pattern with Mandarin bare nouns and that German uncontracted definites pattern with Mandarin demonstratives.²

The parallel-corpus study we will present reveals that this prediction is only partly borne out. We argue that the main problem lies with Mandarin bare nouns, and we

²Strictly speaking, Schwarz does not make claims about Mandarin and Jenks does not make claims about German. This is why the crosslinguistic prediction we look into is attributed to Schwarz's and Jenks's work combined.

develop a more fine-grained typology of strong definiteness. Furthermore, we argue that the predictions *Index!* makes are too strict, even for German.

6.3 Corpus and methodology

As indicated above, the main prediction that follows from Schwarz's and Jenks's work combined is that German contracted definites pattern with Mandarin bare nouns and that German uncontracted definites pattern with Mandarin demonstratives. Parallel corpora allow us to test this prediction. We present our corpus and methodology in Section 6.3.1 and Section 6.3.2, respectively, and provide results and discussion in Section 6.4.

6.3.1 Corpus

The corpus we selected is the first volume of the Harry Potter series and its translations into German and Mandarin (see bibliographic details following the reference list). We chose this corpus for several reasons. First, we want this research to lead to a broader exploration of article systems in languages that are traditionally considered article-less. To do so, we need a corpus that can easily be extended with more languages. The availability of the Harry Potter series in multiple typologically diverse languages provides exactly that. Second, we wanted to have a source language that does not formally distinguish between weak and strong definiteness. This guarantees that the translator is not biased by a formal distinction in the source text and focuses on rendering the meaning in the best way possible. For this reason, we decided to study a corpus with an English source text rather than one with a Mandarin or a German source text.

6.3.2 Methodology

Data selection

We selected our data on the basis of the German translation. We did this because German is known to mark the weak/strong definiteness distinction in a restricted domain—namely, prepositional phrases (PPs). Starting from Mandarin bare nouns and demonstratives or starting from English definites would have led to a high number of data points with no relevant comparative material in German.

The goal of our selection procedure was to end up with a dataset with a more or less even distribution of contracted and uncontracted PPs, at the same time maximiz-

ing the likelihood of including minimal pairs. To do so, we extracted all PPs with a definite article, divided them into contracted and uncontracted ones, and then selected those that contained prepositions that appeared in both lists. A further selection was done for contracted PPs, as these greatly outnumbered uncontracted PPs. In particular, we included all contracted PPs from the first three chapters and restricted the contracted PPs from the other chapters to those that had uncontracted counterparts in the novel—that is, uncontracted PPs involving the same preposition and noun.³ Our selection procedure gave rise to a total of 96 data points, including 40 contracted and 56 uncontracted PPs.

Data processing

Once the set of German PPs was established, we aligned them with the English original and the Mandarin translation. We also annotated all data points for the forms that were used in the three languages. After annotation, each of the 96 data points could be characterized as a triple (German form, English form, Mandarin form) (e.g., ⟨contracted, definite, bare⟩). Alignment and annotation were done by two of the authors, one a native speaker of German, the other a native speaker of Mandarin. All data processing was done in a custom-made online interface that links data and annotation to a number of parallel-corpus analysis tools.

Analysis

We calculated basic descriptive statistics for all data points, including the frequency of forms per language and the frequency of the correspondences between German and Mandarin. Given that the number of data points and languages is limited, this could have sufficed. However, as indicated above, we hope this research will lead to a broader exploration of articles in languages that are traditionally considered article-less. Given that correspondences between more than two languages become difficult to process with basic descriptive statistics, we need another type of analysis that can help us do so. One such family of analyses is known as proximity maps (see Georgakopoulos and Polis, 2018 for discussion). These have gained traction in the typological literature and have been shown to hold promise for cross-linguistic work at the

³The automated scripts we used for extraction and selection are available at <https://github.com/time-in-translation/conll-extractor> and https://github.com/time-in-translation/conll-extractor/blob/master/conll_extractor/prepositions/data.py.

syntax-semantics interface as well (see van der Klis et al., 2020). We introduce them in this article as a proof of concept.

We rely on a specific implementation of proximity maps known as probabilistic semantic maps (Wälchli and Cysouw, 2012). They are a powerful tool for analyzing the use of language-specific forms across data points drawn from a parallel corpus. Here, we provide an intuitive explanation of how probabilistic semantic maps are built and how they can be interpreted.⁴ The example maps we treat are based on our data and will recur in Section 6.4.

Probabilistic semantic maps

Probabilistic semantic maps are generated through Multidimensional Scaling (MDS), a dimensionality reduction algorithm. Each data point is represented by a dot in a two-dimensional space, as in Figure 6.1. All other things being equal, data points are closer to one another if they use the same form in a given language. The more forms that correspond between two data points, the closer they are. We illustrate with the triples in (8).

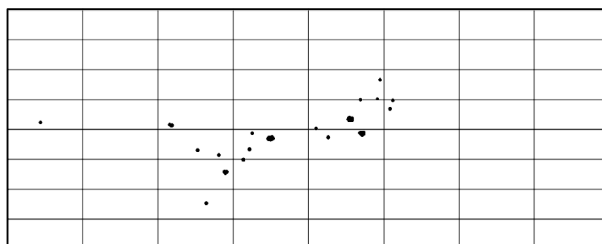


Figure 6.1: A probabilistic semantic map.

(8) *Examples of triples*

- a. ⟨contracted, definite, bare⟩
- b. ⟨contracted, definite, bare⟩
- c. ⟨uncontracted, definite, bare⟩
- d. ⟨uncontracted, definite, demonstrative⟩

All other things being equal, a dot corresponding to a data point like (8a) will be closer to one corresponding to (8b) than to one corresponding to (8c): (8a) and (8b)

⁴For technical details, see Wälchli and Cysouw 2012, van der Klis, Le Bruyn, and de Swart 2017, van der Klis and Tellings 2020.

share all forms, whereas they differ from (8c) in their first position. At the same time, a dot corresponding to a data point like (8d) will be even farther away from (8a)/(8b) than (8c) is, because it differs from (8a)/(8b) in even more forms.

Probabilistic semantic maps that are faithful to all distances between data points are rare. This is because they are limited to two dimensions. MDS can, however, be run with as many dimensions as we like. Dimensions will try to be faithful to the distances between as many data points as possible, but they will progressively also try to be faithful to distances that earlier dimensions were not yet faithful to. This allows us to choose the dimensions that best allow us to study the correspondences (or lack thereof) between forms of different languages.

By default, we run MDS with five dimensions. In Figure 6.1, we have represented the first two. Figure 6.2 shows how we can use these to visualize correspondences between forms. Each shaded cluster in Figure 6.2 represents the distribution of a form across the data points in the corpus. For convenience, we have limited ourselves to two forms from two languages: the clusters with dotted lines represent the two forms from language A; the clusters with solid lines represent the two forms from language B.

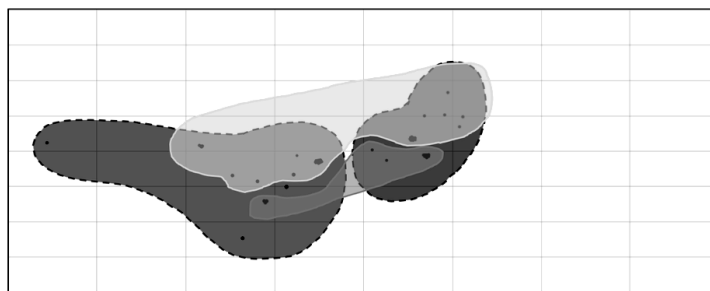


Figure 6.2: A probabilistic semantic map with language-specific form-based clusters.

If there had been a clear correspondence between forms in languages A and B, we would have expected the clusters from the two languages to resemble each other. This is clearly not what happens in Figure 6.2: rather than revealing a similar distribution, the map suggests that the distribution of the forms in the two languages is orthogonal.

6.4 Results and discussion

6.4.1 Results

Descriptive statistics

We indicated above that the German dataset consists of 40 (41.5%) contracted and 56 (58.5%) uncontracted cases. For English and Mandarin, we report on the forms that appeared at least three times as counterparts of one of these. For English, these are the definite ($n = 80$, 83%), the bare singular ($n = 5$, 5%), and the demonstrative ($n = 4$, 4%). For Mandarin, they are the bare noun ($n = 79$, 82%) and the demonstrative ($n = 13$, 13.5%).

As for the correspondences between German and Mandarin, we also restrict ourselves to those forms that appear at least three times as counterparts. Among the 40 German contracted forms, 3 (7.5%) correspond to demonstratives in Mandarin and 34 (85%) correspond to bare nouns. Among the 56 German uncontracted forms, 10 (18%) correspond to demonstratives in Mandarin and 45 (80.5%) correspond to bare nouns.

A probabilistic semantic map

Even though the descriptive statistics already give a good idea of the data, probabilistic semantic maps allow us to visualize them in a format that is easier to process, in particular when more languages are added. Figure 6.3 is identical to Figure 6.2, but the language-specific form-based clusters have now been identified. Crucially, we find that Mandarin bare nouns are the counterparts of both contracted and uncontracted cases in German and that the same holds for Mandarin demonstratives. We thus find that the division of labor between bare nouns and demonstratives is not parallel but orthogonal to the one between contracted and uncontracted definites in German.

6.4.2 Discussion

The data strongly suggest that the main prediction that stems from Schwarz's and Jenks's work combined is not borne out. The one-to-one mappings we expect between German contracted definites and Mandarin bare nouns on the one hand, and between German uncontracted definites and Mandarin demonstratives on the other hand, are not there. A more fine-grained discussion of the data is required, though. We start with

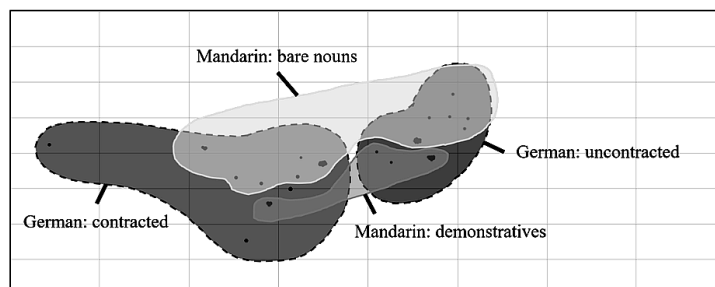


Figure 6.3: A probabilistic semantic map with the main form-based clusters for German and Mandarin.

the German data (Section 6.4.2) and then move to Mandarin demonstratives (Section 6.4.2) and bare nouns (Section 6.4.2). We end with a summary (Section 6.4.2).

German: confirmation of Schwarz’s analysis and problems for Index!

The German data show that unique referents are contracted and that familiar referents are uncontracted. This is in line with the basic predictions of Schwarz’s analysis.

- (9) **E:** I suppose we could take him to the zoo,’ said Aunt Petunia slowly, ‘... and leave him in the car ...
G: Ich denke, wir könnten ihn in den Zoo mitnehmen,’ sagte Tante Petunia langsam, ‘... und ihn im Wagen lassen ...

- (10) [Context: As the owls flooded into the Great Hall as usual, everyone’s attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]
E: They had hardly fluttered out of the way when another owl dropped a letter on top of the parcel.
G: Sie waren kaum aus dem Weg geflattert, als eine andere Eule einen Brief auf das Paket warf.

The *car* in (9) does not refer back to a previously introduced car; rather, it refers to the unique family car. It consequently counts as a weak definite. The *parcel* in (10)

refers back to the package that was introduced before and therefore counts as a strong definite. As Schwarz's analysis predicts, German relies on a contracted definite in (9) and an uncontracted definite in (10).

There are two types of contexts that deserve special mention. Both combine a dimension of uniqueness with a dimension of familiarity. They thus count as in-between cases, and we expect there to be some variation in the markers that appear. The first type is concerned with reference to a familiar but one-of-a-kind stone known as *the Philosopher's Stone*. The second type involves bridging. We find that both contracted and uncontracted definites can be used to refer to the Philosopher's Stone and that bridging is equally variable.

(11) E: 'I'm going out of here tonight and I'm going to try and get **to the Stone** first.'

G: 'Ich gehe heute Nacht raus und versuche als Erster **zum Stein** zu
I go today night out and try as first to.the stone to
kommen.'
get

(12) E: 'How do you think you'd get **to the Stone** without us?'

G: 'Wie glaubst du eigentlich, dass du ohne uns **zu dem Stein**
how believe you actually that you without us to the stone
kommst?'
get

(13) [Context: 'OUT!' roared Uncle Vernon, and he took both Harry and Dudley by the scruffs of their necks and threw them into the hall, slamming the kitchen door behind them.]

E: Harry and Dudley promptly had a furious but silent fight over who would listen **at the keyhole**.

G: Prompt lieferten sich Harry und Dudley einen erbitterten, aber
promptly gave themselves Harry and Dudley a furious but
stummen Kampf darum, wer **am Schlüsselloch** lauschen durfte.
silent fight about who at.the keyhole listen could

(14) [Context: Ducking under Peeves they ran for their lives, right to the end of the corridor, where they slammed into a door—and it was locked.]

E: 'Oh, move over,' Hermione snarled. She grabbed Harry's wand, tapped **the lock** and whispered, 'Alohomora!'

G: ‘Ach, geh mal beiseite,’ fauchte Hermine. Sie packte Harrys
 oh go once away snarled Hermione she took Harry’s
 Zauberstab, klopfte **auf das Türschloss** und flüsterte: ‘Alohomora!’
 wand tapped on the doorlock and whispered alohomora

The Philosopher’s Stone is referred to with a contracted definite in (11) and with an uncontracted definite in (12). Neither of them counts as the first reference to the stone. In (13) and (14), *Schlüsselloch* and *Türschloss* refer to the unique lock of a previously introduced door, but the former appears with a contracted definite whereas the latter combines with an uncontracted definite.

The data in (11)–(14) show that as soon as a dimension of uniqueness is combined with a dimension of familiarity, both contracted and uncontracted definites become available. This is compatible with Schwarz’s basic analysis but undermines the validity of Jenks’s constraint *Index!*. This constraint predicts that the use of weak definites is proscribed as soon as familiarity comes into play. The data in (11) and (13) show that this prediction is not borne out.⁵ We assume that the choice between contracted and uncontracted definites in (11)–(14) is not free, but that the constraints at work go beyond the basic distinction between weak and strong definiteness. For a discussion of some of the factors involved, see Aguilar-Guevara and Zwarts 2010.

We conclude that the German data are in line with the predictions Schwarz (2009) makes and that they argue against the stricter competition between weak and strong definites that follows from Jenks’s constraint *Index!*.

Mandarin: the case of demonstratives

With German following Schwarz’s predictions, we turn to the Mandarin data to understand the orthogonality between German contracted/uncontracted definites and Mandarin bare nouns/demonstratives. In this section, we focus on Mandarin demonstratives.

Our Mandarin dataset contains 13 demonstratives. The vast majority of them ($n = 10$, 77%) appear in contexts that take an uncontracted definite in German and thus behave the way Jenks predicts; that is, they are used in strong definiteness contexts. We argue that the three remaining demonstratives are not to be considered counterexamples to Jenks’s predictions. (15) is a representative example.

⁵Data like those in (13) and (14) are also interesting for a discussion of the more involved claims Schwarz makes about the relation between weak and strong definites in different types of bridging. A reviewer notes that part of the explanation might lie in the fact that *door* is part of *Türschloss* (lit. ‘door lock’) but not of *Schlüsselloch* (lit. ‘keyhole’). A similar effect of compounding is mentioned by Schwarz (2009:283).

(15) E: “I’m not having one **in the house**, Petunia!”

M: “Pèinī, wǒ juébù ràng tāmen rènhéren jìn **zhè dòng fángzi**.”
 Petunia I not have them anyone enter this CL house

(15) is uttered by a husband who assures his wife that certain people will never be welcome in their house. One can argue that the demonstrative is used to refer to the unique family home, but a more plausible analysis is that the demonstrative retains its full deictic force and refers to the house the speaker and listener are in at the moment of speech. Our consultants confirm that the latter analysis is the one that corresponds best to their intuitions. This extends to the two other cases of demonstratives appearing as counterparts of German contracted definites.

We conclude that the Mandarin demonstratives in our corpus mark strong and not weak definiteness. This is in line with Jenks’s analysis. If demonstratives do appear as counterparts of contracted definites, they retain their full deictic force and receive a slightly different interpretation from the one conveyed by the German translation.⁶

Mandarin: the case of bare nouns

Having argued that the few Mandarin demonstratives that do not follow Jenks’s predictions can be considered special cases and do not challenge his analysis, we now turn to Mandarin bare nouns.

Our Mandarin dataset contains 79 bare nouns. They are the majority option, both in contexts in which German uses contracted definites (34 (85%) are rendered as bare nouns) and in contexts in which German uses uncontracted definites (45 (80.5%) are rendered as bare nouns). The use of bare nouns in the former contexts is in line with Jenks’s predictions, but their use in the latter poses a serious challenge. Examples like (16) and (17) show that this challenge is real.

(16) [Context: As the owls flooded into the Great Hall as usual, everyone’s attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

M: tāmen pūshan-zhe chībǎng gānggāng fēi zǒu, yòu yǒu yī zhī
 they flutter-ASP wings right fly away and have one CL

⁶Jenks (2018) proposes a unified analysis of demonstratives that covers both deictic and strong definite uses. We remain neutral as to whether a unified or an ambiguity analysis should be pursued.

māotóuyīng xié lái yī fēng xìn, rēng zài bāoguǒ shàngmiàn.
owl bring come one CL letter throw to parcel on

‘They had hardly fluttered out of the way when another owl dropped a letter on top of the parcel.’

- (17) [Context: Dudley quickly found the largest snake in the place. It could have wrapped its body twice around Uncle Vernon’s car and crushed it into a dustbin—but at the moment it didn’t look in the mood. In fact, it was fast asleep.]

E: He looked back **at the snake** and winked, too.

G: Er drehte sich wieder **zu der Schlange** um und zwinkerte
he turned himself again to the snake around and winked
zurück.
back

M: Tā huí-guò tóu lái kàn-zhe **jù mǎng**, yě duì tā zhǎ-le-zhǎ
he back-ASP head to stare-ASP huge snake too to it wink-ASP-wink
yǎn.
eye

(16) is the Mandarin version of (10). In (16) and (17), reference is made, respectively, to a parcel and a snake that were introduced before. This anaphoric reading is also the reading our consultants find most natural. Examples (16) and (17) thus unambiguously show that bare nouns can be used in strong definiteness contexts.

We conclude that the Mandarin bare nouns in our corpus mark both weak and strong definiteness. These data are incompatible with Jenks’s analysis of the weak/strong opposition in Mandarin.

Summary

Our corpus data show that there is no one-to-one correspondence between German contracted/uncontracted definites and Mandarin bare nouns/demonstratives. This runs counter to the predictions that stem from Schwarz’s and Jenks’s work combined. A closer analysis of the data reveals that German contracted and uncontracted definites, as well as Mandarin demonstratives, by and large behave as expected (Section 6.4.2–Section 6.4.2). The main problem lies with Mandarin bare nouns, as they appear in both weak and strong definiteness contexts (Section 6.4.2). A further issue our data raise is that Jenks’s constraint *Index!* is too strong, not only for the Mandarin data but

also for the German data. In Section 6.5, we reflect on the acceptability of bare nouns in weak and strong definiteness contexts.

6.5 Two types of strong definiteness

The results in Section 6.4 show that Mandarin bare nouns appear in weak and strong definiteness contexts alike. Here, we consider how these facts affect our understanding of definiteness. At the heart of our discussion lies a discrepancy between our data and Jenks's: Jenks uses examples like (4) to argue that bare nouns are ungrammatical as anaphors, whereas we find examples like (16) in which bare nouns are perfectly fine as anaphors. We hypothesize that the opposition indicates two types of strong definiteness: text-level and situation-level familiarity. We start, however, by briefly discussing two competing hypotheses.

6.5.1 Dialectal variation

An obvious hypothesis about the opposing judgments for (4) and (16) is that they stem from dialectal differences. This could be the case, as our consultants are from mainland China whereas Jenks's are from Taiwan. However, our consultants agree with Jenks's that (4) is unacceptable. The opposition between the unacceptability of the bare noun in (4) and its acceptability in (16) is thus real.

6.5.2 Pragmatic coreference

Another hypothesis one could entertain is that the anaphoric reading of (16) is pragmatically induced rather than semantically encoded. Under this hypothesis, the fact that the package in (16) is identified with the package that was introduced before is driven by context and not by semantics. Even though this hypothesis could explain why the bare noun in (16) can have an anaphoric reading, it would need to be supplemented with an explanation for the fact that pragmatic coreference is not an option for the bare noun in (4). We do not see what this explanation could look like; instead, we turn to an alternative hypothesis in which anaphoricity is semantically encoded in both (4) and (16).

6.5.3 Text-level vs. situation-level familiarity

The hypothesis we argue for is based on a comparison of Jenks's data with our corpus data. The contexts in our corpus typically display a classical narrative style in which events are presented in chronological order. (16) is a good example, chronologically relating the coming and going of a group of owls followed by an event involving a single owl. (4) is crucially different in the sense that the event referred to in the second sentence chronologically precedes the one in the first. We hypothesize that the discrepancy in judgments about contexts like those in (4) and (16) is related to this difference in narrative structure and indicates a difference between two types of strong definiteness: text-level and situation-level familiarity. We introduce the two types, make explicit how we take Mandarin bare nouns and demonstratives to relate to them, and discuss the predictions our hypothesis makes.

Two types of familiarity

We start by introducing the two types of familiarity on the basis of the English versions of (4) and (16) (repeated here).

(18) There are a boy and a girl sitting in the classroom. I met **the boy** yesterday.
(= (4))

(19) [Context: As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]
They had hardly fluttered out of the way when another owl dropped a letter on top of **the parcel**. (= (16))

In (18), *the boy* refers back to the previously introduced boy, and in (19), *the parcel* refers back to the previously introduced long thin package. Both meet the requirement of previous introduction traditionally associated with strong definiteness. In Schwarz's and Jenks's analyses of strong definiteness, this requirement is formalized through an identity relation with a pragmatically supplied index. We refer to this type of strong definiteness as *text-level familiarity*.

Situation-level familiarity is stricter and requires the anaphor to be introduced in the same topic situation as its antecedent. We argue that *the parcel* in (19) meets this

requirement but *the boy* in (18) does not. As we indicated in Section 6.2.2, we build on McKenzie (2012, 2015)'s work and take spatiotemporal contiguity to be a good indicator of which eventualities can be considered part of a single situation. In (19), the sentences linking *a long thin package* and *the parcel* describe spatiotemporally contiguous eventualities, and they can thus be assumed to be part of a single overarching topic situation. *The parcel* thus meets the requirements of situation-level familiarity. Example (18) is different: the adverb *yesterday* introduces a clear temporal break between the situations described by the first and second sentences. *The boy* consequently does not meet the requirements of situation-level familiarity.

Bare nouns, demonstratives, and strong definiteness

With the two types of strong definiteness in place, we can present the way we assume Mandarin bare nouns and demonstratives relate to them. We hypothesize that bare nouns can be used for situation-level familiarity but not for text-level familiarity. Demonstratives, on the other hand, can be freely used for both types.

The underlying intuition is that indices are only available in the topic situation in which they have been introduced. The difference between bare nouns and demonstratives lies in the deictic component of the latter: demonstratives are able to refer to situations other than the topic situation of the sentence they appear in and thus to access indices from other topic situations.⁷ A full formalization lies beyond the scope of this article, but the crucial step lies in enriching the analyses Schwarz and Jenks propose for strong definites with a mechanism that allows us to keep track of how the topic situations of different sentences relate to one another. This means we need a dynamic interpretation not only of the indices in (5b) and (6b) but also of the situation pronouns. With this mechanism in place, we can work out the relationship between indices and situation pronouns to derive the difference between text- and situation-level familiarity

The Mandarin versions of (18) and (19) illustrate our hypothesis. *Nánshēng* 'boy' in (4) meets the requirements of text-level familiarity but not those of situation-level familiarity. This explains Jenks's observation that Mandarin requires the demonstrative in this context. *Bāoguǒ* 'parcel' in (16) does meet the requirements of situation-level familiarity, and this explains our finding that Mandarin allows the use of the bare noun in this context.

⁷Wolter (2006) argues that English demonstratives need to be interpreted with respect to nondefault situations. A full comparison between the conditions of use of Mandarin and English demonstratives regrettably lies beyond the scope of this article.

Predictions

If our hypothesis is on the right track, we expect that manipulating the spatiotemporal contiguity of eventualities in examples like (18) and (19) leads to changes in the acceptability of bare nouns. Corpora do not allow us to check the outcomes of these manipulations, so we turn to consultants and their judgment.

Above, we presented (18) as involving two eventualities that are spatiotemporally disjoint. We hypothesized that this is why the bare noun is unacceptable in the Mandarin version. Example (20) is a minimally different variant.

(20) There were a boy and a girl in the classroom. I entered and hit **the boy**.

(20) is different from (18) in that the state of there being a boy and a girl in the classroom and the event of the speaker going in and hitting the boy spatiotemporally overlap and can straightforwardly be thought of as being part of a single overarching topic situation. We thus have set up a context in which we no longer have to rely on text-level familiarity but can also resort to situation-level familiarity. In line with our hypothesis, we predict the demonstrative to remain available but the bare noun to become a viable option as well.

(21) is the Mandarin version of (20):

(21) Jiàoshì lǐ yǒu yí gè nánshēng hé yí gè nǚshēng. Wǒ jìn
 classroom in have one CL boy and one CL girl I enter
 jiàoshì dǎ-le **nánshēng**.
 classroom hit-ASP boy
 ‘There were a boy and a girl in the classroom. I entered and hit **the boy**.’

Our consultants report that a demonstrative can be added to *nánshēng* ‘boy’ in the second sentence of (21) but that this is not required. When asked to compare the acceptability of (21) with the Mandarin version of (18) (i.e., Jenks’s original example), they indicate that there is a clear difference between the two: (21) is acceptable without the demonstrative whereas (18) is not. These judgments are in line with our predictions: if the antecedent is introduced in the same topic situation as its anaphor, the latter can be realized as a bare noun.

Let us turn to (19). Above, we presented it as a context in which all eventualities are part of a single overarching topic situation. In (22), we present a slight modification.

(22) [Context: As the owls flooded into the Great Hall as usual, everyone’s attention was caught at once by a long thin package carried by six screech owls. Harry

was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

M: Mài gé jiàoshòu qián yì tiān jì gěi hā lì #(zhè ge) bāoguǒ.
McGonagall Professor before one day send to Harry this CL package

‘Professor McGonagall had sent the package to Harry the day before.’

(22) is different from the Mandarin version of (19) in that the eventuality of sending the package is spatiotemporally disjoint from all other eventualities in the context. Situation-level familiarity is consequently no longer available. In line with our hypothesis, we predict (22) to differ from the Mandarin version of (19) in that the bare noun is no longer a viable option to mark familiarity. This prediction is borne out, as our consultants report that *bāoguǒ* ‘package’ in (22) requires the demonstrative.

6.5.4 Summary

In this section, we defended the hypothesis that the acceptability of bare nouns in strong definiteness environments in Mandarin indicates that there are two subtypes of strong definiteness: text-level and situation-level familiarity. Demonstratives can mark both, but bare nouns are limited to the latter. We showed how the hypothesis explains our data and generalizes to new contexts.

6.6 Conclusion

Schwarz (2009) and Jenks (2018) argue that the weak/strong definiteness distinction is active in German and Mandarin, respectively. We carried out a parallel-corpus study to check the cross-linguistic predictions that follow. Our corpus data show that the distributions of German contracted/uncontracted definites and Mandarin bare nouns/demonstratives are orthogonal (Section 6.3). Closer scrutiny of the data reveals that the problem lies with Mandarin bare nouns, as they appear in both weak and strong definiteness contexts (Section 6.4). We argued that the discrepancy between our data and Jenks’s indicates that there are two types of strong definiteness: text-level and situation-level familiarity. Bare nouns turn out to be compatible only with situation-level familiarity (Section 6.5).

Our results also shed light on the competition between weak and strong definiteness markers. Jenks’s constraint *Index!* leads to a strict separation between contexts

allowing for weak and strong definites. Setting the Mandarin facts aside, we found that this constraint is too strong, even for our German data.

Our data led us to maintain that German contracted/uncontracted definites are uniformly weak/strong. Mandarin demonstratives also turn out to be uniformly strong but Mandarin bare nouns turn out to be ambiguous between weak and strong definites, the latter being restricted to situation-level familiarity. Another way to go—suggested by a reviewer—is to assume that Mandarin bare nouns uniformly mark weak definiteness. On this analysis, their anaphoric uses would be indicative of the overlap between weak and strong definiteness contexts. We leave the exploration of this competing analysis for future work. The main challenge it faces is to explain why German consistently opts for its strong definite in these contexts whereas Mandarin can also rely on its weak definite.

We conclude with a methodological note. This article has shown the potential of parallel-corpus research for formal approaches to language. On the one hand, we have shown how a small study can lead to relevant results. On the other hand, we have laid the foundation for larger-scale studies, both at the level of corpus compilation and at the level of analysis.

CHAPTER 7

Analyzing the Competition between Bare nouns and Demonstratives

7.1 Introduction

The thesis began with the goal of resolving the alternation challenge: the tension between the functionalist claim that Mandarin bare nouns alternate with demonstratives in definite contexts and with numeral-*yi* in indefinite contexts, and the formal semantic theories that predict no such alternations should exist. As for the definite domain, the initial empirical evidence from Chapter 2 showed that English NPs preceded by *the* are aligned to both Mandarin bare nouns and demonstratives, setting the stage for a deeper inquiry into the division of labor between the two forms in the definite domain. However, in further cross-linguistic comparisons in Chapters 3 and 4, the results failed to confirm the co-existence of bare noun and demonstratives as a systematic alternation unique to Mandarin. Therefore, while the investigation into the indefinite domain concluded with a resolution of the alternation challenge in Chapter 5, the definite domain continues to pose a puzzle.

Chapter 6 tested the most prominent formal explanation for the alternation: the proposal by Jenks (2018) that the division of labor between Mandarin bare nouns and demonstratives reflects the distinction between weak/unique and strong/anaphoric

definiteness, along the lines of German according to Schwarz (2009). Uniqueness of a referent is encoded in German with contracted articles, while anaphoricity is encoded in German with uncontracted articles. Chapter 6 operationalized the weak/strong distinction through a translation corpus study that included German with such an overt morphological distinction between weak (contracted) and strong (uncontracted) definite articles. In this way, we were able to examine the distribution of Mandarin bare nouns and demonstratives in Jenks's weak and strong definite contexts. Our results led to a rejection of the proposal. While demonstratives indeed existed in strong/anaphoric contexts, they were not the only form used in these contexts, and bare nouns occurred freely in both weak/unique and strong/anaphoric definite contexts. This finding undermines Jenks's hypothesis, and compels a re-examination of the division of labor between the two forms.

In this chapter, we put forward what can be considered the null hypothesis: there is no genuine alternation between Mandarin bare nouns and demonstratives in encoding definiteness. Instead, we argue that Mandarin demonstratives are, in fact, canonical anaphoric demonstratives. Bare nouns, in turn, function as the default expression of definiteness. The alternation challenge in the definite domain is resolved if there is no real competition between bare nouns and demonstratives, as each fulfills a distinct semantic role. Bare nouns undergo a covert iota type-shifter, while demonstratives retain their original demonstrative function.

The theoretical challenge, then, shifts to the claim that the observed distribution of these two forms follows predictably from their independent semantic roles within the adapted Properties Approach. To formally test this new hypothesis, we adopt the framework by Ahn (2022), who draws a precise semantic distinction between anaphoric definites and anaphoric demonstratives. Within this framework, we can derive testable predictions about the conditions that favor one form over the other. These predictions offer a new perspective on the division of labor between demonstratives and bare nouns in anaphoric contexts—one that aligns with the empirical observations of Chapter 6, but without equating demonstratives with definite articles. In the rest of this chapter, we return to the data and present a follow-up corpus study of anaphoric bare nouns and demonstratives in the Mandarin translation of *Harry Potter and the Philosopher's Stone* designed to answer the following question:

Does the empirical distribution of Mandarin bare nouns and demonstratives in strong/anaphoric definite contexts align with the formal distinction between anaphoric definites and anaphoric demonstratives as defined

by Ahn (2022), thereby supporting a theoretical analysis of demonstratives as demonstrative markers rather than definite articles?

This chapter is structured as follows. Section 7.2 presents our integration of Ahn’s (2022) analysis into the type-shifting framework of the Properties Approach and derives testable predictions. Section 7.3 introduces the corpus study designed to test these predictions, and Section 7.4 presents and discusses the results, arguing that Mandarin demonstratives appear precisely in those contexts where anaphoric demonstratives are expected. Our findings will be contextualized against recent experimental work on the topic, in particular Saha et al. (2024), who argue that Mandarin prefers demonstratives across anaphoric contexts. In Section 7.5, we conclude that Mandarin demonstratives are run-of-the-mill anaphoric demonstratives and do not function as definite articles, thereby resolving the alternation challenge for the definite domain.

7.2 Anaphoric definites and demonstratives in Ahn (2022)

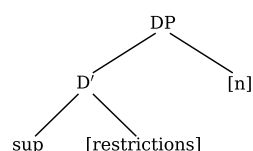
In this section, we present Ahn (2022)’s analysis of definites and demonstratives with a focus on their anaphoric variants (Section 7.2.1), translate it to the type-shifting framework we rely on in this dissertation (Section 7.2.2), and discuss the predictions we make for the distribution of anaphoric definites and demonstratives (Section 7.2.3).

7.2.1 Ahn (2022)

Ahn (2022) hypothesizes that the difference between demonstratives and definites lies in the type of maximality operator they are associated with: definites are associated with a unary maximality operator whereas demonstratives come with a binary one.

For an anaphoric definite like *the man*, Ahn proposes the syntax in (1) and the derivation in (2):

- (1) The syntax of anaphoric definites



(2) The semantic derivation of anaphoric definites

$$\llbracket \text{sup} \rrbracket = \lambda P. \iota x : \forall y [P(y) \leftrightarrow y \leq x]$$

$$\llbracket \text{restrictions} \rrbracket = \lambda x [\text{man}(x)]$$

$$\llbracket [n] \rrbracket = \llbracket \text{id}_{x_1} \rrbracket(n) = \lambda x : x = g(n).x$$

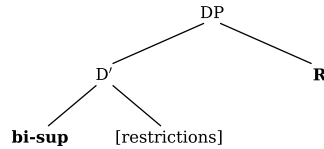
- a. $\lambda P. \iota x : \forall y [P(y) \leftrightarrow y \leq x] \quad \lambda x [\text{man}(x)]$
- b. $\iota x : \forall y [\text{man}(y) \leftrightarrow y \leq x]$
- c. $\lambda x : x = g(n).x \quad \iota x : \forall y [\text{man}(y) \leftrightarrow y \leq x]$
- d. $\iota x : x = g(n). \forall y [\text{man}(y) \leftrightarrow y \leq x]$

Sup is a maximality operator that combines with a noun like *man*, asserts that there is a unique man, and outputs the individual corresponding to that man (2a, 2b). Anaphoricity comes in separately through $[n]$. The core of $[n]$ is the complex function id_{x_1} that takes an index n and returns the function that takes an individual and returns that same individual with the presupposition that it is identical to n . When we combine the result of $\llbracket \text{sup man} \rrbracket$ with $[n]$, we obtain the unique individual that is a man and is presupposed to be identical to the index n (2c and 2d).

For demonstratives, Ahn proposes that their binary maximality operator takes two properties. The first corresponds to the nominal predicate the demonstrative combines with. For the second property, Ahn foresees three options: (i) the (singleton) set of individuals occurring at the location the speaker points at, (ii) the (singleton) set of individuals corresponding to a silent index, and (iii) the set of individuals identified by an overtly spelled out relative clause. If the demonstrative combines with option (i), the semantic effect is that of direct reference akin to that of the (deictic) referential view on demonstratives (Kaplan, 1989). When combining with options (ii) or (iii), the semantic effect is similar to that of the indirectly referential view on demonstratives (e.g., King, 2001; Elbourne, 2008). Ahn's analysis of demonstratives can thus be said to offer a middle-ground view between directly referential and indirectly referential analyses of demonstratives. Our focus here lies on the anaphoric demonstrative (option (ii)).

The syntax Ahn proposes for an anaphoric demonstrative like *that man* is given in (3), its semantic derivation in (4):

(3) The syntax of anaphoric demonstratives



(4) The semantics of anaphoric demonstratives

$$\llbracket \text{bi-sup} \rrbracket = \lambda P \lambda R . \iota x : \forall y [P(y) \& R(y) \leftrightarrow y \leq x]$$

$$\mathbf{R} = \llbracket \text{id}_{x_2} \rrbracket (n) = \lambda x . x = g(n)$$

$$[\text{restriction}] = \lambda x [\text{man}(x)]$$

a. $\lambda P \lambda R . \iota x : \forall y [P(y) \& R(y) \leftrightarrow y \leq x] \quad \lambda x [\text{man}(x)]$

b. $\lambda R . \iota x : \forall y [\text{man}(y) \& R(y) \leftrightarrow y \leq x] \quad \lambda x . x = g(n)$

c. $\iota x : \forall y [\text{man}(y) \& y = g(n) \leftrightarrow y \leq x]$

Bi-sup is a new type of maximality operator that takes two properties (*P* and *R*). The semantic import of *P* is not different from the one we find for the anaphoric definite in (2) but the way anaphoricity comes into play is. $[n]$ in (2) involves the function id_{x_1} that is different from the function id_{x_2} that we find in (4): whereas the former takes an individual and returns that same individual with the *presupposition* that it is identical to the index *n*, the latter takes an individual and returns the *assertion* that it is identical to the index *n*. The end result of the derivation in (4) is one in which the anaphoric demonstrative refers to an individual whose uniqueness is said to be related to the semantic content of the noun it combines with and on its identity to the index *n*. This is crucially different from the end result in (2), where anaphoricity ends up being presupposed and uniqueness is said to be related solely to the semantic content of the noun *Sup* combines with.

Ahn does not develop an extensive set of predictions about what the opposition between (1)/(2) and (3)/(4) means for the competition between anaphoric definites and anaphoric demonstratives. She suffices with two more general considerations. The first is that the *bi-sup* operator is more complex than the *sup* operator, which she takes to lead to a generalized preference for definites over demonstratives. The second is that definites can refer uniquely or anaphorically whereas demonstratives that are not accompanied by a pointing gesture or a relative clause necessarily receive an anaphoric interpretation. As such, demonstratives—unlike definites—allow speakers

to unambiguously signal anaphoricity. Ahn assumes that this property of demonstratives counteracts the otherwise generalized preference for definites.

7.2.2 A type-shifting implementation of Ahn (2022)

While maintaining the semantic differences Ahn hypothesizes, we redefine the semantics of anaphoric definites and demonstratives in (5) and (6), aligning them with the single entries for type-shifters we are familiar with in the type-shifting framework:

- (5) The semantics of anaphoric definites:

$$\llbracket \textit{definite}_{\textit{anaphoric}} \rrbracket = \lambda P \lambda n. \iota x : x = g(n). \forall y [P(y) \leftrightarrow y \leq x]$$

- (6) The semantics of anaphoric demonstratives:

$$\llbracket \textit{demonstrative}_{\textit{anaphoric}} \rrbracket = \lambda P \lambda n. \iota x : \forall y [P(y) \& y = g(n) \leftrightarrow y = x]$$

The different positions of the equation to $g(n)$ in (5) and (6) reflect the two differences Ahn captures with the opposition between the unary and binary maximality operators in (2) and (4). The first difference is that anaphoricity is presupposed for the definite and asserted for the demonstrative. The second difference is related to the status of uniqueness. Even though both the definite and demonstrative refer uniquely, only the definite requires P to correspond to a singleton set.

We take the meanings in (5) and (6) to be standard type-shifts from the domain of predicates to the domain of individuals. Following Ahn for English, (5) is realized as *the* and demonstratives are the overt spell out of (6). For Mandarin, the question we tackle in this chapter is whether the demonstrative blocks the bare noun from undergoing both (5) and (6) or only from undergoing (6).

7.2.3 Predictions

To determine whether Mandarin demonstratives only function as anaphoric demonstratives or have extended their use and also function as anaphoric definites, we need to work out the predictions the analyses in (5) and (6) make about the distribution of bare nouns and demonstratives in Mandarin. The general prediction Ahn makes is too coarse-grained for these purposes and it is also unclear how we could mimic it in a type-shifting framework. Indeed, the prediction Ahn proposes is that there is a generalized preference for anaphoric definites because of their simpler maximality operator. We could make a similar claim about the type-shift in (5) being simpler than the one in (6), but in a type-shifting framework, this would mean that the shift in (6) would

never be available, effectively blocking the use of anaphoric demonstratives across the board. This is clearly not what we want. Instead, we want to take seriously the two semantic dimensions (5) and (6) differ in – anaphoricity and uniqueness – and work out the predictions that come with these.

For anaphoricity, we noted that – on the analyses in (5) and (6) – definites *pre-suppose* anaphoric reference whereas demonstratives *assert* it. We take this to mean that a speaker/writer using an anaphoric definite has every reason to assume that its anaphoric nature is obvious to the hearer/reader whereas the use of an anaphoric demonstrative imposes itself when the anaphoric nature is not obvious.

What it means for the anaphoric nature of an expression to be obvious is likely to involve a set of factors with different languages being sensitive to different subsets of these. The literature on salience (among others, Ariel, 1991; Gundel et al., 1993; Grosz et al., 1995; Almor, 1999; Kaiser, 2003) provides us with a number of possible factors to look into. There are two that are generally agreed on. The first is distance: under the assumption that the activation of a referent in discourse decays over time, the bigger the distance between an intended anaphor and its antecedent, the less obvious the anaphoric nature of the anaphor will be. The second factor is grammatical function: the referents of subjects are assumed to have a higher activation than those of non-subjects. We consequently expect the anaphoric nature of expressions that have non-subjects as their antecedents to be less obvious. Given that we focus on different types of nominal anaphora, we think it is also relevant to add a third factor, inspired by the work of Almor (1999), namely, variations in the nominal content of anaphors. One of the insights that Almor proposes is that the decay over time we referred to before can be counterbalanced by using the same nominal content for the antecedent and its anaphor. The upshot of this is that the use of different nominal content can have a negative impact on the obviousness of the anaphoric nature of an expression. Next to distance and grammatical function, we will consequently also look at changes in nominal content as a relevant factor to track the difference between the presupposed/assertoric nature of anaphoric reference.

Moving to uniqueness, the difference between definite and demonstrative anaphoricity is that the former but not the latter requires the nominal part of an anaphoric expression to refer to a singleton. For anaphoric demonstratives, this means that they can be used in any context in which there are multiple referents that satisfy their descriptive content. Anaphoric definites are different and are only acceptable in those contexts in which there is a single referent satisfying their descriptive content.

Distance, nominal variation and non-uniqueness are the factors we will pay at-

tention to in our data analysis. If Mandarin demonstratives function as full-fledged definite articles, we expect these factors not to have an influence on the distribution of bare nouns and demonstratives in anaphoric contexts. If Mandarin demonstratives do not function as full-fledged definite articles, we expect these factors to provide a neat division between those contexts in which we find demonstratives and those in which we find bare nouns.

7.3 Introducing our study

7.3.1 Corpus

Our corpus consists of all the chapters of the Mandarin translation of *Harry Potter and the Philosopher's Stone*. Testing environments in papers on anaphoric properties are in general limited to first mention referential expressions and their first anaphoric pickup. To replicate these conditions as closely as possible, we selected all occurrences of bare nouns and demonstratives that are the translation of anaphoric definites and are the first nominal anaphoric pickup of referents introduced by $yi+CL+N$ in the Mandarin translation ($n = 64$).

7.3.2 Annotation

To explore the differences between demonstratives and bare nouns, we added annotations that allow us to track the differences at the level of anaphoricity and uniqueness discussed in Section 7.2.3. These are concerned with distance, grammatical function, nominal change and uniqueness. Below, we go through the operationalization of each of them.

Distance

We added two distance-based annotations. The first is *adjacency*: those anaphors that occur in the same sentence as their antecedent or in the sentence immediately following that of their antecedent are annotated as *adjacent*. All others are annotated as *non-adjacent*.

Given the fairly high number of *non-adjacent* anaphors ($n = 52$), we looked for a way to formally distinguish ‘bigger’ from ‘smaller’ distances. Corpus data like (7) suggested a straightforward way of doing so:

- (7) He pulled out the Cloak and then his eyes fell on the flute Hagrid had given him for Christmas.

(7) is taken from Chapter 16. At this point in the story, it is unlikely for the flute that Hagrid gave to Harry in Chapter 12 to still be sufficiently activated for the reader to make the connection between the flute in (7) and Hagrid's Christmas gift. We assume this is why the author added the relative clause 'Hagrid had given him for Christmas' and why the translator followed suit. Extended descriptions that are clearly added to remind the readers about where they first read about the referent led us to add *distance* as an annotation. Anaphors with an extended description are annotated as *far*. All others are annotated as *close*. The number of *far* anaphors is fairly low ($n = 3$). We submit, though, that this subset of anaphors is particularly insightful. Distance is relevant only insofar as it operationalizes decay in activation levels and this is exactly what *far* anaphors are about: with extended descriptions, the author herself indicates that she considers the distance between the antecedents and the anaphors too big for the activation levels of referents to still allow for simple anaphoric pick up.

Grammatical function

We added one annotation for grammatical function. In line with the generally higher activation of grammatical subjects, we added the annotation *subjecthood*, anaphors with subject antecedents receiving the value *subject*, all other anaphors receiving the value *non-subject*.

Nominal change

For nominal change, we added one annotation, *viz.*, *nominal.change*. To determine the value for this annotation, we primarily focused on nouns and checked whether they were maintained from the antecedent to the anaphor where possible. In a number of cases, we also looked into modifiers and classifiers. We discuss the relevant types of decisions we made on the basis of (8) to (12):

Antecedent	Anaphor
(8) <i>no change</i>	
yì fēng xìn one CL letter 'a letter'	xìn letter 'the letter'

(9) *change*

yí gè rén
one CL person

‘a person’

zhè ge xiǎo lǎotóu
that CL small old_man

‘that small old man’

(10) *no change*

yí gè fēicháng qíguài de mèng
one CL very strange DE dream

‘a very strange dream’

zhè ge mèng
that CL dream

‘that dream’

(11) *change*

yí gè dài dōumào de shēnyǐng
one CL wear hood DE figure

‘a figure wearing a hood’

nà ge chuān zhe dǒupéng de
that CL wear ASP cape DE
shēnyǐng
figure

‘that cape wearing figure’

(12) *change*

yí juàn yángpí zhǐ
one CL sheepskin paper

‘a roll of parchment’

nà zhāng zhǐ
that CL paper

‘that paper’

(8) and (9) are straightforward, the former illustrating a case in which the noun is maintained, annotated as *no change*, the latter a case of nominal change, annotated as *change*. (10) and (11) illustrate how we dealt with modifiers. If a modifier is not carried over from the antecedent to the anaphor as in (10) but the noun itself is maintained, we considered this not to be instance of nominal change. However, if the modifier changes as in (11), we did consider this a case of nominal change. Finally, in cases like (12), we considered that the classifier change has an important influence on the way the referent corresponding to the noun is conceptualized: *juàn* is used for rolled up objects, *zhāng* for flat objects. We consequently classified the antecedent-anaphor pair in (12) as involving nominal change.

Uniqueness

For uniqueness, we added a single annotation, *viz.* *uniqueness*. In case the context allowed us to determine that there were multiple referents satisfying the descriptive

content of the anaphor, we set the value to *non-unique*. In all other cases, we set the value to *unique*.

7.3.3 Statistical analysis

To help us explore the tendencies we found in our data, we rely on conditional inference trees (CIT, Tagliamonte and Baayen, 2012). The output of CIT organizes dimensions of variation as a decision tree. What the model does is to check whether the independent variables we selected—the factors we annotated for—are significantly associated with the response variable in the dataset—the choice between a bare noun and a demonstrative. If so, it evaluates which of them has the strongest association and uses the outcome of this evaluation to introduce a binary split in the dataset based on the values of the independent variable. These steps are repeated until no further significant associations are found ($\alpha = 0.05$).

7.4 Results and discussion

As we indicated in Section 7.2.3, sensitivity of demonstratives to the factors we annotated for argues in favor of an analysis that takes them to be run-of-the-mill demonstratives, in line with the semantics in (6). Lack of sensitivity argues in favor of an analysis that takes them to cover both the semantics in (6) and the one in (5), doing double duty as demonstratives and as full-fledged anaphoric definite articles.

The graph in Figure 7.1 presents the output of the conditional inference tree analysis (CIT) we ran. Its tree shape allows for easy visual inspection: the oval nodes (1, 3, and 4) represent factors that have a significant effect and the branches indicate the splits in our dataset, leading either to other oval nodes or to bin nodes (2, 5, 6, and 7). The latter present the division in distribution of bare nouns and demonstratives for a relevant subset of our data (e.g., node 2 presents the distribution of bare nouns and demonstratives for all anaphors involving nominal change).

The CIT reveals that the distribution of demonstratives is sensitive to three of the factors we annotated for: demonstratives ($n = 24$) are outnumbered by bare nouns ($n = 40$) and are the majority option only for anaphors that are non-unique, are far away from their antecedents, or involve nominal change. Bare nouns are the majority option for all anaphors that do not involve nominal change, are close to their antecedents and count as unique. As such, the CIT supports an analysis of demonstratives in which they do not double as full-fledged anaphoric definite articles.

In this section, we provide discussions for the individual oval nodes of the CIT (Section 7.4.1–Section 7.4.3) and then move to a general discussion (Section 7.4.4). In the latter, we weigh our results, reflect on the demonstrative’s lack of sensitivity to adjacency and subjecthood, on the difference in the use of demonstratives in the Mandarin translation and the English original, and on the implications of our results for the experimental data of Saha et al. (2024).

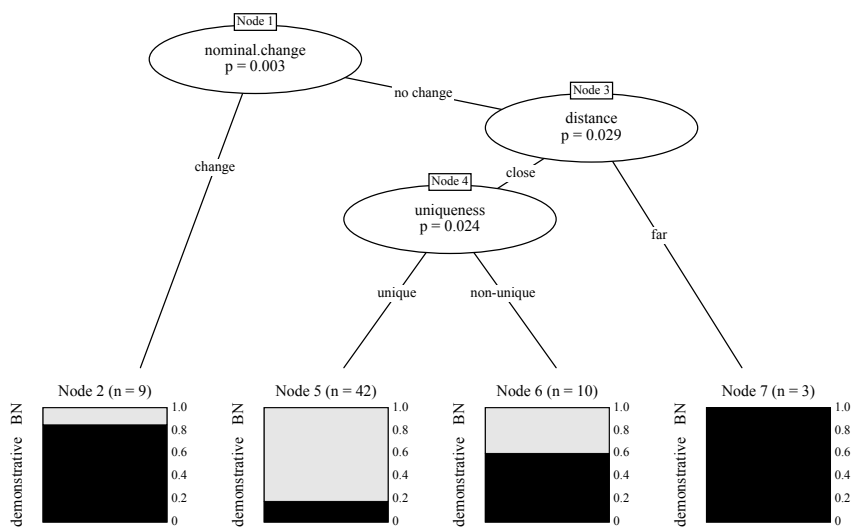


Figure 7.1: Conditional Inference Tree output for the distribution of bare nouns and demonstratives in all chapters of *Harry Potter and the Philosopher’s Stone*

7.4.1 Node 1: nominal change

In node 1 of the CIT, we find that `nominal.change` is the factor with the highest predictive value for the full dataset. We remind the reader that we expected demonstratives to be preferred with anaphors with a nominal change in view of the fact that nominal changes can blur anaphoric connections and that the demonstrative helps the speaker/writer to make these connections explicit again. Our expectation is borne out: `nominal.change` comes out as a significant factor and we find a demonstrative in 7 out of the 8 cases involving nominal change. We illustrate with (13), and also briefly look at one of the bare noun cases in (14). For easy reference, we provide the context from the English original and the sentences with the relevant anaphors and antecedents

in Mandarin.

- (13) [Context: He found it a lot harder to concentrate on drills that afternoon, and when he left the building at five o'clock, he was still so worried that he walked straight into someone just outside the door.

'Sorry,' he grunted, as the tiny old man stumbled and almost fell.]

Antecedent sentence:

yǔ zhàn zài ménkǒu de **yí gè rén** zhuàng le gè mǎnhuái
with stand at doorway DE a CL person bump ASP CL fully

'He bumped right into a person standing in the doorway.'

Anaphor sentence:

zhè ge xiǎo lǎotóu dǎ le gè lièqiè
this CL small old man make ASP CL stumble

'This little old man stumbled'

The example in (13) illustrates how the change from *rén* ('person') to *xiǎo lǎotóu* ('small old man') leads to the use of a demonstrative in Mandarin. We interpret the use of the demonstrative here as a way to make explicit that *xiǎo lǎotóu* refers to a previously introduced individual.

- (14) [Context: A giant of a man was standing in the doorway. His face was almost completely hidden by a long, shaggy mane of hair and a wild, tangled beard, but you could make out his eyes, glinting like black beetles under all the hair. The giant squeezed his way into the hut, stooping so that his head just brushed the ceiling.]

Antecedent sentence:

ménkǒu zhàn zhe **yí gè biāoxíng dàhàn**.
doorway stand ASP a CL tiger.build big man

'A burly man was standing at the entrance.'

Anaphor sentence:

jùrén hǎo bù róngyì cái jǐ jìn wū lái
giant very not easy just squeeze enter house come

'The giant squeezed into the house with great difficulty'

For (14), we observe that there is a nominal change: from *biāoxíng dàhàn* ('a big burly man', lit. 'tiger.build big man') to *jùrén* ('giant'). Even though the use of the bare noun in this example might consequently come as a surprise, we think this should be qualified. Indeed, *biāoxíng dàhàn* and *jùrén* are strictly speaking different nouns, but they highlight exactly the same characteristics of their referent, making it extremely unlikely that a demonstrative would be needed to make the anaphoric connection between the two explicit.

We conclude that Mandarin demonstratives are used in the context of nominal change and that this is connected to their role of making explicit anaphoric connections that would be blurred in their absence. This role of demonstratives is also central to the factor *distance* that we turn to next.

7.4.2 Node 3: distance

Within the set of anaphors not involving nominal change, *distance* appears as the next significant factor in our CIT. We remind the reader that this factor distinguishes between anaphors that come with an extensive modifier reminding the reader of the moment their referents were first introduced and anaphors that do not come with this type of modifier. Our interpretation of the distinction is that anaphors that come with this type of modifier are such that they are too far away from their antecedent to allow for simple anaphoric pickup. We consequently annotated anaphors with the relevant type of modifier as *far* and those without it as *close*. Under the assumption that demonstratives help the speaker/writer to make anaphoric connections explicit, we expect them to pop up with *far* anaphors. What the CIT analysis shows is that this is indeed the case: all three *far* anaphors come with a demonstrative. We illustrate with the example in (15):

(15) Context:

[Chapter 12] Harry picked up the top parcel. It was wrapped in thick brown paper and scrawled across it was To Harry, from Hagrid. Inside was a roughly cut wooden flute. Hagrid had obviously whittled it himself. Harry blew it—it sounded a bit like an owl.

[Chapter 16] 'Better get the Cloak,' Ron muttered, as Lee Jordan finally left, stretching and yawning. Harry ran upstairs to their dark dormitory. He pulled out the Cloak and then his eyes fell on the flute Hagrid had given him for Christmas. He pocketed it to use on Fluffy—he didn't feel much like singing.

Antecedent sentence:

lǐmiàn shì yì zhī zuògōng hěn cūcāo de dízi
inside is a CL craftsmanship very rough DE flute

‘Inside, there was a flute with very rough craftsmanship.’

Anaphor sentence:

tā wúyìjiān kànjiàn le shèngdànjié shí hǎigé sònggěi tā de nà
he inadvertently see ASP Christmas when Hagrid give him DE that
zhī dízi
CL flute

‘He by accident saw the flute Hagrid gave him for Christmas.’

The example in (15) is about a flute that Hagrid gave to Harry for Christmas in Chapter 12 and it plays an important role again in Chapter 16. We submit that it is the big distance between the two references to the flute that pushes the author in the original to add the modifier ‘Hagrid had given him for Christmas’. In Mandarin, this type of modifier is accompanied by a demonstrative.

We conclude that Mandarin demonstratives are used when the anaphor and antecedent are far apart and that this is connected to their role of making explicit anaphoric connections that would be blurred in their absence.

7.4.3 Node 4: uniqueness

Within the set of anaphors that do not involve nominal change and are qualified as close, the final factor that the CIT reveals to have a significant effect is *uniqueness*. As the reader can verify in the CIT, there is a clear contrast between anaphors that we classified as unique and the ones we classified as non-unique, the former having a distinct preference for bare nouns, the latter a distinct preference for demonstratives. At the same time, these preferences are less pronounced than for nodes 1 and 2. In our discussion, we consequently look into four types of anaphors, crossing unique/non-unique anaphors and bare noun/demonstrative ones.

Unique anaphors with a bare noun

The example in (16) illustrates a case of a unique anaphor that is rendered with a bare noun:

- (16) [Context: He sat up and Hagrid’s heavy coat fell off him. The hut was full of sunlight, the storm was over, Hagrid himself was asleep on the collapsed sofa

and there was an owl rapping its claw on the window, a newspaper held in its beak.

Harry scrambled to his feet, so happy he felt as though a large balloon was swelling inside him. He went straight to the window and jerked it open. The owl swooped in and dropped the newspaper on top of Hagrid, who didn't wake up.]

Antecedent sentence:

yì zhī māotóuyīng zhèng yòng zhuǎzi qiāodǎ chuānghù
one CL owl PRT using claws knock window

'An owl was knocking on the window with its claws.'

Anaphor sentence:

māotóuyīng fēi le jīnlái
owl fly ASP enter

'The owl flew in.'

The context in (16) is taken from Chapter: Harry and Hagrid wake up and an owl arrives to bring the newspaper. The owl is unique in this setting and is picked up with a bare noun (*māotóuyīng*).

If demonstratives were doing double duty as full-fledged definite articles and would be the standard way of rendering the definite semantics in (5), we would expect bare nouns to be outnumbered by demonstratives for unique anaphors. This is not what we find: out of the 42 cases of unique anaphors in the CIT, 35 are rendered as a bare noun and only 7 as a demonstrative. This strongly suggests that bare nouns are still the default way of rendering the semantics of the definite article in (5). One question that remains to be answered, though, is what triggers the presence of demonstratives in the small subset of unique anaphors in which we find them. This is the question we turn to next.

Unique anaphors with a demonstrative

We find a number of cases in which a unique anaphor is rendered with a demonstrative. In view of the fact that the large majority of unique anaphors takes a bare noun, we consider these cases exceptional and we take them to add a meaning dimension that goes beyond the English original. These cases, we argue, involve an affective use of the demonstrative or, as in (17), a spatial one:

- (17) [Context: We received your message and enclose your Christmas present. From Uncle Vernon and Aunt Petunia. Sellotaped to the note was a fifty-pence piece.

‘That’s friendly,’ said Harry.

Ron was fascinated by the fifty pence.]

Antecedent sentence:

yòng tòumíng jiāodài zhān zài zhǐ tiáo shàng de shì yì méi wǔshí
using transparent tape stick on paper note on DE is one CL fifty
biànshì de yìngbì.
pence DE coin

‘Attached to the note with scotch tape was a fifty-pence coin.’

Anaphor sentence:

Luóēn bèi nà méi yìngbì mí zhù le.
Ron BEI that CL coin captivated ASP

‘Ron was captivated by that coin.’

The example in (17) involves Ron and Harry going over their Christmas presents. Even though they are sitting in the same room, there is a distance between them. We submit that the translator chose Harry’s perspective when reporting on the reading of the small note from his aunt and uncle, but that it is Ron’s perspective that is chosen to report his fascination. On this analysis, the use of the demonstrative for the anaphor in (17) is a standard spatial (distal) use.

Affective and spatial uses squarely fall within the boundaries of run-of-the-mill uses of demonstratives but do not fall under Ahn’s analysis of anaphoric demonstratives. We consequently take the exceptional cases in which we find demonstratives to render unique anaphors not to justify an analysis in which demonstratives are taken to cover part of the semantics of a regular definite article in (5).

Non-unique anaphors with a bare noun

Where unique anaphors typically take a bare noun and demonstratives are the exception, our data show that non-unique anaphors typically take demonstratives and that bare nouns are the exception. Rather than first looking at the typical case of a non-unique anaphor rendered as a demonstrative, we think it is instructive to look at the exception, viz. non-unique anaphors rendered as bare nouns:

- (18) [Context: A pair of goblins bowed them through the silver doors and they were in a vast marble hall. About a hundred more goblins were sitting on high stools behind a long counter, scribbling in large ledgers, weighing coins on brass scales, examining precious stones through eyeglasses. There were too many doors to count leading off the hall, and yet more goblins were showing people in and out of these. Hagrid and Harry made for the counter. ‘Morning,’ said Hagrid to a free goblin. ‘We’ve come ter take some money outta Mr Harry Potter’s safe.’ ‘You have his key, sir?’ ‘Got it here somewhere,’ said Hagrid and he started emptying his pockets on to the counter, scattering a handful of mouldy dog-biscuits over the goblin’s book of numbers. The goblin wrinkled his nose.]

Antecedent sentence:

“zǎo,” Hǎigé duì yí gè xiánzhe de yāojing shuō
 morning Hagrid to one CL free DE goblin say

“‘Morning,’ Hagrid said to a free goblin.”

Anaphor sentence:

yāojing zhòu le zhòu bízǐ.
 goblin wrinkle LE wrinkle nose

‘The goblin wrinkled his nose.’

The English excerpt in (18) introduces multiple goblins: the pair of goblins bowing to Harry and Hagrid, the hundred more goblins sitting behind a long counter and the free goblin that Hagrid decides to talk to. There is thus no singleton corresponding to *yāojing* and the anaphor *yāojing* counts as non-unique. The fact that we do find a bare noun, we submit, can be explained if we assume that there is a straightforward way of restricting the context in which the bare noun anaphor appears: even though there are lots of goblins in the context, there is only one that Hagrid is engaged with in an exchange. Given that this exchange is such that the goblin’s reactions follow naturally from Hagrid’s actions, it is likely that the intended referent corresponds to this particular goblin. We conclude that non-uniqueness need not lead to the use of a demonstrative if the context allows us to single out a unique referent.

Non-unique anaphors with a demonstrative

With the exception of non-unique anaphors rendered with a bare noun in place, we can look at the typical case of a non-unique anaphor rendered with a demonstrative:

- (19) [Context: He had to start somewhere. Setting the lamp down carefully on the floor, he looked along the bottom shelf for an interesting-looking book. A large black and silver volume caught his eye. He pulled it out with difficulty, because it was very heavy, and, balancing it on his knee, let it fall open. A piercing, blood-curdling shriek split the silence—the book was screaming!]

Antecedent sentence:

tā tūrán kànjiàn yì běn hēisè hé yínsè xiàng jiǎn de dà shū.
he suddenly see one CL black and silver alternating DE big book

‘He suddenly saw a large black and silver book.’

Anaphor sentence:

nà běn shū zài cǎn jiào
that CL book ASP scream

‘that book was screaming’

The example in (19) is taken from Chapter 12 and is part of Harry’s visit to the Hogwarts library. There is an abundance of books in the library and the noun *shū* therefore does not have a corresponding singleton. The crucial difference with a context like (18), we submit, is that there is no clear reason to assume that the screaming of the book is a reaction to Harry opening it whereas the wrinkling of the goblin’s nose in (18) is a clear reaction to Hagrid’s actions. As such, there is no clear way in which one can restrict the books involved in the screaming to the book Harry took off the shelf. *Shū* thus qualifies as a run-of-the-mill non-unique anaphor and, as the CIT in (7.1) shows, these take a demonstrative in our data.

7.4.4 General discussion

Our discussion of the different oval nodes of the CIT in Section 7.4.1 to Section 7.4.3 confirms the general picture we introduced at the beginning of this section: the distribution of demonstratives in our corpus is sensitive to three of the factors that Ahn’s analysis predicts to govern the distinction between full-fledged anaphoric definites and anaphoric demonstratives. As such, our data provide strong support for the claim that

Mandarin demonstratives function as demonstratives, aligning with the semantics in (6), and that their use does not extend to that of full-fledged anaphoric definites on the analysis in (5). Mandarin demonstratives are thus used to assert anaphoricity and to deal with non-uniqueness.

Despite the fact that our data are clearcut enough, there remain a number of questions that we need to address:

- (i) If demonstratives are sensitive to the distinction between assertion/presupposition, why is it we have not found an effect for *adjacency* and *subjecthood*? In Section 7.2 and 7.3, we argued that these factors also probe the assertion/presupposition distinction.
- (ii) If demonstratives in Mandarin are full-fledged demonstratives along the lines in (6), why is it that we find them as translations of English definites? Does this state of affairs not strongly suggest that Mandarin demonstratives double as anaphoric definites along the lines in (5)?
- (iii) How is it possible that we find that the demonstrative is the minority option in Mandarin? These findings are in direct contradiction to the recent experimental findings of Saha et al. (2024).

For the missing effect of *adjacency* and *subjecthood*, we argue that this state of affairs is probably due to the type of anaphors we studied and our annotation decisions. For *adjacency*, we noted earlier that the number of adjacent anaphors is low. This is likely due to the fact that bare nouns and demonstratives compete with pronouns for this type of anaphors. A more fine-grained measure for relatively small distances might consequently have revealed an effect, but we opted for an extra measure that maximized the impact of big distances instead. With a low number of adjacent anaphors, it is also reasonable to expect *subjecthood* to have a smaller impact: the bigger the distance, the lower we expect the importance of the grammatical function of the antecedent to be. We conclude that there is no reason to see the presence of an effect for *nominal.change* and *distance* to be in conflict with the absence of an effect for *adjacency* and *subjecthood*: both pairs of factors tap into slightly different aspects of salience and can consequently lead to different effects depending on the corpus and annotation decisions.

Moving to the second question, we understand that our claim that Mandarin demonstratives are run-of-the-mill demonstratives may seem controversial in view of the fact that they appear in about a third of the translations of English anaphoric definites. We

remind the reader that the argumentation in favor of our claim is based on the fact that we only find demonstratives in those cases in which we expect to find them: those in which anaphoricity needs to be asserted to highlight anaphoric connections and those in which we are dealing with non-uniqueness. Our second question should then be interpreted as whether a unified analysis of demonstratives can allow for slightly different distributions in Mandarin and English. We argue for an affirmative answer and build it up on the basis of a discussion of the different factors we found Mandarin demonstratives to be sensitive to.

For *nominal.change*, we estimate that Mandarin demonstratives are used differently from English demonstratives: the use of a demonstrative in (13) feels marked in English whereas we consider it marked or even ungrammatical to leave it out in Mandarin. However, *nominal.change* is a dimension of salience and—as we pointed out in Section 7.2.1—we do not have reason to assume that every language is sensitive to the same dimensions of salience. The fact that we estimate that there is a difference in the use of English and Mandarin demonstratives in contexts like (13) consequently does not mean that the semantics of demonstratives in the two languages is fundamentally different.

For the factor *distance*, we also estimate that there is a difference in the use of English and Mandarin demonstratives. In (15), the use of a demonstrative in English would feel marked at the very least whereas it feels perfectly natural in Mandarin. In the same way as for *nominal.change*, we do not think that differences in use because of *distance* endanger a cross-linguistically unified analysis of demonstratives. *Distance* is another salience-based factor that can play out differently across languages and this is arguably independent of the semantics of demonstratives.

For *nominal.change* and *distance*, we have argued that they are salience-based factors and that cross-linguistic variation in the way salience plays out does not endanger a unified analysis of demonstratives. *Uniqueness* is a fundamentally semantic notion and the variation we find in non-unique anaphoric contexts like (19) might consequently seem like a bigger threat to a unified analysis of demonstratives: the use of a demonstrative feels obligatory in Mandarin for the non-unique anaphor in (19) and this leads one to expect the demonstrative to be obligatory in English as well, contrary to fact. We argue that the difference in the way the two languages deal with demonstratives in non-unique anaphoric contexts does not rule out a unified analysis. The key to understanding why a cross-linguistically stable semantic trait like non-uniqueness need not lead to an identical cross-linguistic distribution lies in the existence of cases like (18). The opposition between (18) and (19) shows that the choice

for the bare noun in (18) hinges on discourse factors: if the translator is confident that the reader will single out a unique referent based on the surrounding discourse, a bare noun is an option for non-unique anaphors in Mandarin in the same way as the definite is an option in English. We take this to suggest that the real difference between English and Mandarin in (19) is not that they have a different semantics for the demonstrative, but rather that Mandarin is less liberal in allowing for contextual restrictions and that this leads to a higher frequency of demonstratives in non-unique anaphoric contexts.

Based on the discussion of the different factors that we found to influence the distribution of bare nouns and demonstratives, we conclude that their influence does not reveal profound differences between Mandarin and English demonstratives. We link differences in the distribution of the latter to differences in the way salience and contextual restrictions play out in the two languages. The fact that we find demonstratives as translations of definites consequently does not lead us to question the cross-linguistic validity of the semantics for demonstratives in (6) nor for our claim that Mandarin demonstratives are not doing double duty as anaphoric definites along the lines of the semantics in (5).

The final question we turn to is how it is possible that our results are different from Saha et al. (2024). In Chapter 6, we discussed Jenks (2018) take on Mandarin demonstratives as anaphoric definite article-like expressions. The core of Jenks's argument is based on the obligatory use of a demonstrative in the Mandarin translation of *the boy* in (20):

(20) There are a boy and a girl sitting in the classroom. I met **the boy** yesterday.

The rationale is as follows: if the demonstrative did not play some kind of blocking role akin to that of an anaphoric definite article, we would expect the bare noun to be acceptable as the translation of *the boy*, contrary to fact. In replies to Jenks (2018), we and Dayal and Jiang (2022) counter that contexts like (20) are not neutral testing cases for articlehood as there is a spatial and temporal break between the situation described in the first sentence and the event described in the second sentence. Both we and Dayal and Jiang (2022) further provide attested examples of contexts in which the break is absent and the bare noun occurs (see also Simpson and Wu, 2022).

In 2024, Saha et al. joined the discussion and added an experimental perspective, arguing that the spatial and temporal break in (20) cannot be held responsible for the reduced acceptability of the bare noun translation of *the boy*. (21) and (22) present the crucial opposition from their experimental paradigm. The discourse in (21) replicates the break we find in (20), the discourse in (22) does not present such a break.

- (21) A boy and a girl entered the classroom. I had noticed **the boy** at a coffee shop yesterday.
- (22) A boy and a girl entered the classroom. **The boy** sat down in the front row.

The cross-linguistic results of Saha et al. (2024) show that the Mandarin translations with demonstratives are considered more acceptable than those with bare nouns for both (21) and (22), whereas other languages prefer other definite expressions over demonstratives (definite article in English, bare nouns in Turkish and noun-classifier combinations in Bangla). Saha et al. (2024) conclude that Mandarin prefers demonstratives in all anaphoric contexts and that the dispreferred status of the bare noun in (20) is independent of the spatial/temporal break. According to these authors, Mandarin demonstratives thus count as anaphoric definite articles, in line with (Jenks, 2018).

What Saha et al. (2024) show is that Mandarin demonstratives have a wider distribution than demonstratives in other languages. At the same time, a full-fledged definite article analysis would be unsatisfactory in view of the anaphoric definite contexts in which we do find bare nouns in Mandarin (see Chapter 5, Dayal and Jiang, 2022; Simpson and Wu, 2022). In the dataset discussed in Chapter 5, bare nouns are even the majority option, making an anaphoric definite article analysis even less desirable.

One way of approaching the difference in results would be to capitalize on methodology: whereas we rely on translated corpus data, Saha et al. rely on experiments involving native speaker judgements. We maintain that both methodologies lead to reliable results and submit that a methodological debate is less likely to shed light on where the differences come from than an analysis of the actual contexts of Saha et al.'s experiment.

We provided the core example of the experiment with the antecedent and the relevant anaphor sentences in Mandarin as below.

- (23) Experimental item Saha et al. (2024)

Antecedent sentence:

yí gè nánhái hé yí gè nǚhái zǒujìn le jiàoshì.
one CL boy and one CL girl walk_into ASP classroom

'A boy and a girl walked into the classroom.'

Anaphor sentence without break:

nà ge nánhái / nánhái zuò zài qiánpái.
 that CL boy / boy sit LOC front_seat

‘**The boy** sat at the front.’

Anaphor sentence with break:

wǒ zuótiān zài shūdiàn jiàn guo **nánhái / nà ge nánhái**.
 I yesterday at bookstore see ASP boy / that CL boy

‘I saw **the boy** yesterday at the bookstore.’

We remind the reader that Mandarin native speakers preferred the anaphor sentences with the demonstrative over those with the bare noun, independently of whether the sentence was designed to have a break. We argue that these results do not come as a surprise and do not show that there is a general preference for anaphoric demonstratives in Mandarin.

We start by noting that (23) is set in a classroom context. We assume that there are typically multiple boys in a classroom and consequently take it that we are dealing with a non-unique anaphor. From our discussion of the use of non-unique anaphors with bare nouns and demonstratives in (18) and (19), we know that non-unique anaphors can take bare nouns but that there should be a straightforward way of linking the anaphor to its antecedent. We also know that the discourse conditions under which this linking is allowed are stricter in Mandarin than in English. For the anaphor sentence with a spatio-temporal break, it is reasonable to assume that the relevant link between the antecedent and the anaphor cannot be established without the demonstrative. We argue that the same holds for the anaphor sentence that Saha et al. designed not to have a break. In the English version in (22), the event of entering the classroom and that of sitting in the front can get a sequential interpretation, allowing a straightforward link between the boy in the antecedent and the anaphor sentences. However, no such sequential interpretation is available in the Mandarin version in (23). The bottom line then is that the anaphor sentence that should not have had a break ended up having one and that the crucial experimental manipulation failed, inevitably leading to a general preference for the demonstrative. We conclude that our results do not contradict the findings of Saha et al. (2024) but that their conclusion that Mandarin has a general preference for demonstrative anaphors is ill-guided and should be attributed to problems in the operationalization of their items.

7.5 Conclusion

This chapter's primary goal was to resolve the definite part of the alternation challenge as to addressing the division of labor between bare nouns and demonstratives. After Chapter 6 disproved Jenks's (2018) weak/strong definiteness hypothesis, this chapter tested the hypothesis that the Mandarin demonstratives in strong/anaphoric definite contexts are not functioning as definite articles but are, in fact, run-of-the-mill demonstratives operating with their anaphoric roles.

To empirically test this hypothesis, this chapter applied the formal analytical framework of Ahn (2022) to quantitative corpus data. This framework provides clear criteria for distinguishing between anaphoric definites and anaphoric demonstratives, centering on whether anaphoricity is asserted versus presupposed and whether the uniqueness condition on the nominal phrase is met. Using a translation corpus from *Harry Potter and the Philosopher's Stone* in its English originals and Mandarin translations, this chapter analyzed Mandarin bare nouns and demonstratives that translate English anaphoric definites. The selection of these cases was constrained to those where the referent's antecedent had been introduced into the Mandarin discourse using the numeral-*yi* construction. The corpus analysis shows that the distribution of Mandarin bare nouns and demonstratives is not random but follows a principled division of labor as predicted by Ahn's proposal: bare nouns are the default choice when anaphoricity can be presupposed and uniqueness is satisfied; demonstratives are deployed when anaphoricity must be explicitly asserted (e.g., due to a change in the referent or distance from its last mention) or when the uniqueness condition is not met.

This empirical finding thus resolves the alternation challenge in the definite domain. Since demonstratives have their own distinct semantics and are not definite articles, the Blocking Principle is not violated. The path for bare nouns to become definite arguments via a covert iota type-shifter remains available, serving as the default choice for encoding definiteness in Mandarin.

CHAPTER 8

Conclusion

8.1 Thesis Goal

The goal of this thesis is to provide a comprehensive answer to the question, "How does a language like Mandarin, which lacks a dedicated article system, systematically encode definiteness and indefiniteness?" While Mandarin bare nouns can be interpreted with flexibility as both definite and indefinite expressions, previous functionalist explorations on other alternation forms in definite and indefinite domains, *viz.*, demonstratives and numeral-*yi* (one), lead to what we term the alternation challenge, the central puzzle of this thesis.

As outlined in Chapter 1, this alternation challenge arises from a tension between functionalist observations and two formalist frameworks on argument formation. According to the functionalist literature, different forms co-occur with bare nouns in both Mandarin definite and indefinite contexts. Bare nouns alternate with NPs preceded by numeral-*yi* constructions (*yi* + classifier) in indefinite contexts and with NPs preceded by demonstrative constructions (*zhè/nà* + classifier) in definite contexts. The functionalist literature proposes that the numeral and demonstrative meanings are bleached out when these forms emerge in indefinite and definite contexts (Chen, 2003, 2004; Li and Thompson, 1989; Lü, 1947; Wright and Givón, 1987).

The formalist literature predicts that no such alternations exist. The formalist semantic theories on argument formation are represented by the two dominant frameworks: the Kinds Approach (Chierchia, 1998) and the Properties Approach (Krifka, 2003). Both approaches view Mandarin bare nouns as self-sufficient in (in)definite contexts, and as no additional support is needed, the alternation with demonstratives and numeral-*yi* is unexpected. Both frameworks are grounded in Partee's (1987) type-shifting mechanism and the Blocking Principle, but they confront the alternation challenge in fundamentally different ways. This divergence arises from distinct theoretical assumptions about how Mandarin bare nouns acquire definite and indefinite interpretations.

Within the Properties Approach, Mandarin bare nouns start their lives as properties of type $\langle e, t \rangle$, parallel to nouns in other languages. They are posited to acquire definite interpretations via the iota (ι) type-shifter and indefinite interpretations via the existential quantifier (\exists). The alternation challenge poses a significant problem for this analysis due to the Blocking Principle. This is because the Blocking Principle dictates that the availability of overt markers, such as demonstratives for definiteness or the numeral-*yi* for indefiniteness, should block the covert application of corresponding type-shifters (ι and \exists). As for the Kinds Approach, Mandarin is analyzed as a prototypical language whose nouns start their lives as kinds of type $\langle e \rangle$. In the most recent formulation by Jiang (2020), Mandarin bare nouns derive definite and indefinite readings through two primary mechanisms: Situation Restriction (SR) and Derived Kinds Predication (DKP). On the one hand, SR maps the kind denoted by a bare noun to its maximal member of that kind within a given situation, yielding a definite reading. On the other hand, DKP converts a sortal predicate with a kind argument into the same predicate with existential quantification over instances of that kind, resulting in an indefinite reading. The alternation challenge is problematic for the Kinds Approach, which struggles to explain the motivation for the existence of overt alternations in the first place. If Mandarin bare nouns are argumental by nature and can readily acquire both definite and indefinite interpretations, there is no reason to develop other explicit expressions to convey the same meanings.

Though their technical details differ, both the Properties Approach and the Kinds Approach predict the absence of an alternation between Mandarin bare nouns and other forms in regular definite and indefinite contexts. Demonstratives and numeral-*yi* are expected to occur in demonstrative and numeral contexts only.

To answer our primary question of how Mandarin encodes definiteness and indefiniteness, we must resolve the alternation challenge. Therefore, this thesis pursues

a step-wise investigation that operationalizes this challenge by breaking it down into three interconnected perspectives.

The first and most fundamental step is to establish the empirical facts and determine whether the alternations described by functionalists occur systematically in the Mandarin referential domain. This foundational inquiry leads to the following empirical question: Do we find only bare nouns, or do we also find numeral-*yi* and/or demonstratives in indefinite and definite contexts? If both are present, what is the division of labor?

Once these empirical patterns are established, the central task then becomes to reconcile this reality with the existing formal semantic framework and to determine what our empirical findings mean for formal semantic approaches to argument formation. We must determine whether we need to develop a fundamentally different alternative to existing approaches like the Kinds and Properties Approaches, or whether we can account for the alternations and derive the division of labor within at least one of these approaches.

Answering the empirical question in a language like Mandarin presents its own methodological challenge. The primary challenge we confront is that without a dedicated article system to overtly mark (in)definiteness, it is difficult to systematically identify these definite and indefinite contexts. This thesis uses a translation corpus methodology as a heuristic strategy. We rely on the English articles *the* and *a/an* as semantic proxies for regular definite and indefinite contexts, and we use *Harry Potter and the Philosopher's Stone* and its translations (henceforth, the HP corpus) as the source of corpus data. The alignment of the Mandarin translation with the English original maps out with the distribution of expressions Mandarin uses in definite and indefinite contexts. The translation corpus methodology constitutes a crucial step in the research, but it depends on the assumption that a translation corpus approach provides reliable data for answering our empirical and, consequently, our theoretical questions. Throughout the investigation, the Mandarin data are placed in a broader multilingual perspective, given that we check the Mandarin patterns against the distribution of bare nouns in other articleless languages with the help of the multilingual translation corpus of the Harry Potter novel.

The structure of this conclusion chapter centers around these three dimensions of inquiry: empirical, theoretical, and methodological. We address each of these dimensions in turn to develop the argumentation across chapters to provide comprehensive answers to our overarching question of how Mandarin encodes definiteness and indefiniteness. Section 8.2 synthesizes the empirical findings. Section 8.3 presents the

adapted theoretical framework to explain the empirical results. Section 8.4 details the methodological innovations. Section 8.5 presents the limitations and future directions. Section 8.6 offers the conclusions.

Chapters 3, 4, and 6 of this thesis are based on content that was previously published as co-authored articles (Chapter 3: *'Articleless' Languages are Not Created Equal*, *Sinn und Bedeutung* 27; Chapter 4: *The Theory of Argument Formation: between Kinds and Properties*, *SALT* 33; Chapter 6: *Translation Mining: Definiteness across Languages. A Reply to Jenks (2018)*, *Linguistic Inquiry*). Each article was originally designed for its respective venue. In this thesis, they have been integrated to serve the general line of argumentation.

8.2 Empirical findings

This section synthesizes the empirical investigation, which proceeded in several stages to answer the guiding empirical question of the thesis posed in Chapter 1:

Empirically, how does Mandarin encode definiteness and indefiniteness?
Do we only find bare nouns or also numeral-*yi* and/or demonstratives in indefinite and definite contexts? If we find both, what is the division of labor?

The investigation began in Chapter 2 with an exploratory analysis of the entire referential system ($n = 1210$) in the first chapter of *Harry Potter and the Philosopher's Stone* and its Mandarin translation. The quantitative results revealed a clear distribution pattern: English NPs preceded by *an* were translated by Mandarin bare nouns in 26% of cases and by Mandarin numeral-*yi* in 65% of cases. Meanwhile, English NPs preceded by *the* were translated by Mandarin bare nouns in 81% of cases and by Mandarin demonstratives in 14% of cases. Furthermore, the analysis of the broader referential system showed that Mandarin demonstratives and numeral-*yi* retained their canonical demonstrative and numeral contexts, appearing as counterparts to English demonstratives (77%) and numerals. We therefore confirm that the alternation patterns claimed by functionalist observations do exist in our translation corpus. The findings immediately raised a critical question: were these patterns a genuine feature of the Mandarin referential system or merely an artifact of the English source text?

To eliminate this possible corpus bias and validate the Mandarin pattern, Chapter 3 moved from a comparison between English original and Mandarin translations

to a cross-linguistic comparison of the same set of English regular definite and indefinite contexts in their Mandarin, Russian and Hindi translations. Russian and Hindi, similar to Mandarin, are two other languages without dedicated definite and indefinite articles, henceforth so-called “article-less” languages (Dayal, 2004). A Fisher’s Exact Test revealed significant differences in indefinite domains across all three languages and showed that Mandarin’s usage of numeral-*yi* was significantly higher than the usage of numeral-one in Russian (for which bare nouns dominated, representing more than 80% of cases). Similarly, Hindi numeral-*yi* (40%) was more frequent than Russian numeral-one. We concluded that the high frequency of Mandarin numeral-*yi* in indefinite contexts found in Chapter 2 is a genuine, language specific feature of Mandarin, validating the alternations in the indefinite domain as a robust empirical phenomenon. For definiteness, while the use of demonstratives in Mandarin was higher than in Russian, the statistical differences were not as significant across all three languages. Before reviewing the results in detail in indefinite and definite contexts, Chapter 4 solidified these observed patterns in previous chapters by expanding the HP corpus to include two languages with well-documented article systems (Spanish and German) and one with only a definite article (Hebrew) to complement the grammatical set-up of the article system. The results aligned perfectly with the existing literature.

After confirming the cross-linguistic robustness of the alternation pattern, we explored how the alternation forms are distributed in (in)definite contexts. Chapter 5 established the division of labor between bare nouns and numeral-*yi* in Mandarin indefinite contexts. The chapter tested the new hypothesis, namely that Mandarin bare nouns in indefinite singular contexts are restricted in a systematic way that corresponds to the intuition of Huang (2015), who argues for a typicality restriction on a limited set of verb-noun combinations involving pseudo-incorporation. The hypothesis was tested on a new, targeted dataset derived from the HP corpus, consisting of English NPs preceded by *alan* in object position and their Mandarin translations as bare nouns and numeral-*yi* ($n = 154$). The results showed that bare nouns were indeed restricted to combinations with verbs where such a typicality relation holds, supporting a new lexical-semantic hallmark of Mandarin pseudo-incorporation. These findings solidified the division of labor between numeral-*yi* and bare nouns in indefinite contexts: bare nouns used for indefiniteness are restricted to pseudo-incorporation characterized by the typicality relation between bare nouns and verbs they combined with, while numeral-*yi* functions as the marker for regular indefinite arguments.

We then investigated the definite domain. Chapter 6 empirically evaluated the alternation between bare nouns and demonstratives following Jenks (2018)’s proposal,

which posits that the alternation between bare nouns and demonstratives reflects the distinctions between the weak/unique and strong/anaphoric definiteness, in line with the contracted and uncontracted articles in German. The hypothesis was tested on a targeted dataset derived from the HP corpus, comprising German and Mandarin translations. We compared the German distribution of contracted ($n = 40$) and uncontracted ($n = 56$) NPs in prepositional phrases, representing weak and strong definites, with the Mandarin distribution of bare nouns and demonstratives in corresponding alignments. The results falsified Jenks (2018)'s proposal on the division of labor between Mandarin bare nouns and demonstratives. Specifically, although Mandarin demonstratives were largely used in strong definite contexts, they were not the only (or even the most frequent) option, as bare nouns also occurred freely in both weak and strong definite contexts. This finding is incompatible with a clean division of labor between the two forms within definite contexts as proposed by Jenks (2018).

The failure of Jenks' hypothesis to account for the data in Chapter 6 led us to test an alternative hypothesis in Chapter 7, namely that no alternation exists in Mandarin definite domain at all. The goal was to test the proposal, in line with Ahn (2022), that Mandarin demonstratives are not definite articles but canonical anaphoric demonstratives. We zoomed in on strong definite contexts, where we observed in Chapter 6 that both bare nouns and demonstratives appeared while their division of labor was uncertain. We conducted the analysis on a targeted dataset derived from the HP corpus with Mandarin translations, consisting of all strong definites in *Harry Potter and the Philosopher's Stone* whose first nominal anaphoric referent was previously introduced by numeral-*yi* constructions ($n = 64$). The results revealed a clear division of labor based on anaphoric function, not definiteness: Mandarin demonstratives were the preferred choice when the anaphoric link was potentially ambiguous or needs to be explicitly asserted, while bare nouns were the default choice for obvious and unambiguous anaphoric references. This confirmed that demonstratives operate in a separate functional domain not as definite articles but as standard demonstratives, while bare nouns remain the primary choice of expressing anaphoric definiteness.

Based on the above empirical investigation, we can provide comprehensive answers to the empirical questions posed at the beginning of this section. The data in our translation corpus show that Mandarin encodes definiteness and indefiniteness through a hybrid system in that we not only find bare nouns but also numeral-*yi* in indefinite contexts and demonstratives in definite contexts. In the indefinite domain, numeral-*yi* functions as a regular indefinite article, while bare nouns are restricted to pseudo-incorporation constructions characterized by the typicality relation with the

verb they combine with. In the definite domain, bare nouns are the default expression; demonstratives are not the definite article but function as regular demonstratives, operating in a separate space for anaphoricity rather than definiteness. These nuanced, complex empirical realities force existing formal theories to account for the alternation challenge, as shown in the next subsection.

8.3 Theoretical contributions

Now that the empirical picture of the Mandarin referential system is sketched in Section 8.2, this section addresses the central theoretical goal of the thesis. That is, it determines what the findings mean for formal semantic approaches to argument formation. We repeat the overarching theoretical question outlined in Chapter 1 below:

Theoretically, what do our empirical findings mean for formal semantic approaches to argument formation? Do we need to develop a fundamentally different alternative to existing approaches like the Kinds and the Properties Approach or can we account for the alternations and derive the division of labor within at least one of these approaches?

As expressed in Chapter 1, the core theoretical puzzle is motivated by the alternation challenge, which originates from the tension between the functionalist claim that bare nouns alternate with numeral-*yi* in indefinite contexts and with demonstratives in definite contexts, along with the predictions of both formal frameworks that such alternations should not exist. For the Properties Approach, this challenge is rooted in the Blocking Principle, which posits that an overt marker (like demonstratives and numeral-*yi*) should block a bare noun from acquiring the same definite and indefinite meanings via a covert type-shift. For the Kinds Approach, the challenge is a lack of motivation: if Mandarin bare nouns are argumental by nature, there is no pressure for the language to develop article-like expressions.

Chapter 4 confronted this alternation challenge. First, it determined which theoretical framework was better equipped to explain the cross-linguistic variations encoding definiteness and indefiniteness. This theoretical decision was driven by the empirical findings from Chapters 3 and 4. Chapter 3 compared Mandarin pattern to other “article-less” languages (Russian and Hindi). Chapter 4 extended the translation corpus of Chapter 3 to include languages with established article system: two with a dedicated article system (Spanish and German), and one with a definite article system (Hebrew). Based on the cross-linguistic comparison, we claimed that the

Properties Approach offered greater flexibility to account for the distributional patterns. Therefore, we adopted the core claims of the Properties Approach that bare nouns start their lives as properties ($\text{type}\langle e, t \rangle$) that undergo covert type-shifting and that they are subject to the Blocking Principle. We proposed several modifications to resolve the obstacle inherent in the original Properties Approach for the alternation challenge, particularly concerning the Blocking Principle. Each modification follows directly from the empirical findings. In the indefinite domain, the data in Chapter 2 revealed that Mandarin's use of numeral-*yi* in indefinite contexts was the dominant expression aligned with English NPs preceded by *a*. Moreover, its use was significantly higher than numeral-one usage in cross-linguistic comparison to Russian and Hindi in Chapters 3 and 4. Thus, we recognized Mandarin numeral-*yi* as an indefinite article. If numeral-*yi* is proposed as an indefinite article, it should block bare nouns from being indefinite in regular argument position via a covert existential type-shifter. In our proposal, bare nouns in indefinite contexts are not regular arguments but are pseudo-incorporated in non-argument positions, thus avoiding the blocking effect of numeral-*yi*. In the definite domain, the data in Chapters 3 and 4 showed no cross-linguistic distinction among Russian, Hindi, and Mandarin, nor did it reveal a clear, systematic alternation pattern. This led us to treat the status of Mandarin demonstratives with caution, acknowledging their uses in definite contexts without committing to their analysis as definite articles.

Chapter 5 develops the theoretical solution for the indefinite domain in detail, and shows that indefinite bare nouns are restricted to verb-noun combinations with a typicality restriction. Based on the empirical finding, we identified this semantic criterion as the hallmark of pseudo-incorporation in Mandarin. Our analysis based on combined insights from previous work on the typicality restriction in verb-noun combinations involving pseudo-incorporation in Mandarin (Huang, 2015) and semantic analyses on pseudo-incorporation (Farkas and Swart, 2003; Espinal and McNally, 2011; Dayal, 2011; Luo, 2022) solves the challenge caused by the fact that Mandarin lacks morpho-syntactic markers to identify the pseudo-incorporation. Following Huang (2015), the typicality restriction on the verb-noun combination in Mandarin pseudo-incorporation instantiates a semantic dependency by which the verb either denotes the typical use of the object or describes the typical event through which the object comes into existence. Our central insight is that the typicality restriction rests on the verb-noun combinations, parallel to Dayal (2011)'s analysis on Hindi. Following Le Bruyn et al. (2016), the typicality restriction on the verb-noun combination is captured through the interaction between the verb's argument structure and the noun's QUALIA roles

(Pustejovsky, 1995). Specifically, the pseudo-incorporation arises when the lexical semantics of the verb doubles the relational information in the noun's QUALIA structure. Crucially, by deriving the existential force compositionally from the verb in the verb-noun combination, the analysis eliminates the need for a covert existential type-shift for bare nouns for indefinite interpretations. As a result, Mandarin bare nouns and numeral-*yi* in indefinite contexts follow distinct derivational paths. Numeral-*yi* encodes indefiniteness in regular argument positions, while bare nouns express indefiniteness in non-argument positions. As the Blocking Principle is not operative in non-argument positions for pseudo-incorporation constructions, the indefinite part of the alternation challenge is resolved.

After accounting for the indefinite alternation, we turned to the definite domain, where the status of alternation between bare nouns and demonstratives was still empirically and theoretically unclear. Chapter 6 began by testing the hypothesis that a genuine alternation—specifically, the weak/strong definiteness distinction proposed by Jenks (2018), which aligns bare nouns with weak definites and demonstratives with strong definites—exists in definite contexts. Our parallel corpus study comparing Mandarin with German definites did not support this hypothesis. While Mandarin demonstratives were mostly applied in strong definite contexts, bare nouns appeared freely in both weak and strong contexts, which is incompatible with the notion that there is a clean division of labor of the two forms in definite contexts. Thus, we rejected Jenks's proposal concerning the division of labor between bare nouns and demonstratives as weak and strong definiteness.

This result motivated the investigation in Chapter 7, which pursued the alternative hypothesis that no genuine alternation exists because Mandarin demonstratives do not function as definite articles but as regular anaphoric demonstratives. We adopted Ahn's (2022) formal framework, which distinguishes anaphoric definites from anaphoric demonstratives based on key semantic differences. For definites, anaphoricity is presupposed and the referent must be unique; for demonstratives, anaphoricity is asserted, and non-uniqueness is tolerated. Our corpus analysis confirmed this distinction, showing that demonstratives are systematically used when anaphoric links are non-obvious due to factors like nominal change or distance or when the referent is not unique. This confirmed that demonstratives operate in a separate domain from definites. Consequently, the alternation challenge in the definite domain is resolved, as bare nouns are the default definite expression and can freely undergo the covert iota type-shift without being blocked. Demonstratives are not competing with bare nouns for definiteness, but they are regular anaphoric demonstratives.

In response to the core theoretical question, this thesis demonstrates that there is no need to develop a fundamentally different alternative to existing formal semantic approaches. Instead, the alternation challenge can be fully resolved within an adapted Properties Approach. Recall that the alternation challenge on Properties Approaches was related to the Blocking Principle, which prevents bare nouns from deriving the same semantic meanings via covert type-shifters when an overt marker like demonstratives and numeral-*yi* is available.

In the indefinite domain, the alternation challenge is resolved by the connections between the Blocking Principle and argumenthood. The numeral-*yi* functions as a genuine indefinite article, which, per the Blocking Principle, still prevents any covert type-shifting operations via bleached existential quantifier deriving an indefinite meaning. Indefinite bare nouns are restricted to pseudo-incorporation constructions, where they receive their existential import from the verb-noun combination, rather than the existential type-shifter.

In the definite domain, the alternation challenge is resolved by demonstrating that Mandarin demonstratives are not definite articles but are regular anaphoric demonstratives, which operate in a separate functional domain. Therefore, they do not compete with bare nouns, leaving bare nouns free to acquire definiteness through the covert iota type-shift.

Thus, by adapting the existing Properties Approach based on empirical facts, we were able to resolve the alternation challenge without abandoning the approach. This theoretical resolution, however, was only possible because the translation approach first provided the empirical clarity needed to distinguish genuine alternation patterns. This brings us to the methodological innovations of the thesis.

8.4 Methodological innovations

The argumentation of this thesis is based on a translation methodology that is a means of data collection and constitutes an innovative approach to the alternation challenge for a language like Mandarin. Since Mandarin lacks a dedicated article system, there is no systematic way to identify definite and indefinite contexts.

The innovation of this thesis lies in the use of a multilingual translation corpus to overcome this obstacle. The starting point resides in a shared theoretical assumption of the Properties and Kinds Approaches, namely, that the English definite article *the* and indefinite article *a/an* are the overt realizations of the iota (ι) and existential (\exists) type-shifters, respectively, and therefore, can serve as semantic proxies for regular

definite and indefinite contexts. This design allowed us to observe the forms occurring in relevant semantic contexts and quantitatively analyze their distributions.

This methodology operated in two distinct but complementary modes throughout the thesis, showcasing its adaptability as a research tool in the referential domain. First, it served as an exploratory tool to map the empirical distributional patterns of how Mandarin referential system encodes (in)definiteness. By examining the Mandarin referential system as a whole, Chapter 2 provided quantitative evidence that bare nouns and demonstrative/numeral-*yi* constructions co-occur in both definite and indefinite contexts while maintaining their canonical functions. The analysis was based on a comparison between the English originals and Mandarin translation of *Harry Potter and the Philosopher's Stone*. The results revealed that numeral-*yi* and demonstratives have a significant extended use in (in)definite contexts alongside their canonical functions. This initial quantitative picture was validated and refined in Chapters 3 and 4. Chapter 3 addressed the methodological question of whether these patterns were genuine features of Mandarin referential system or the artifact of corpus bias in the selection of English source texts. We compared the Russian and Hindi translations of the same English dataset of Chapter 2 and found that the Mandarin pattern is strikingly different from Russian and Hindi patterns, particularly in the indefinite domain (the distinction in the definite domain was not significant). Chapter 4 solidified these findings by adding an extensive layer of comparison that included languages with well-documented article systems like Spanish, German, and Hebrew. We compared the translations of the same English corpus and found that the translation patterns in these languages aligned perfectly with the description in their respective literature. This result provided strong evidence that definiteness and indefiniteness remain stable across translations, thus securing the empirical foundation of the thesis.

Second, the translation corpus functions as a hypothesis-testing tool that allowed us to verify or falsify specific theoretical claims through highly targeted corpora. This function is illustrated in the corpus's applications in Chapters 4 – 7. In Chapter 4, the extended cross-linguistic filter licensed the core theoretical move that the Kinds Approach cannot account for cross-linguistic pattern, while an adapted version of the Properties Approach can. In Chapter 5, we constructed a dataset of all English indefinite objects and their Mandarin translation corpus from the HP corpus. This allowed us to empirically test the proposed lexical-semantic criterion of characterizing pseudo-incorporation in Mandarin. Namely, we tested whether bare nouns in Mandarin indefinite contexts are restricted to combinations with verbs for which the typicality relation holds. In Chapter 6, we applied the cross-linguistic translation corpus

to test Jenks (2018) hypothesis on Mandarin bare nouns and demonstrative division as weak/strong definiteness aligned to German contracted/uncontracted definites. Using the German translation of the HP corpus, we created a specific targeted dataset that compares German contracted and uncontracted definites to their Mandarin counterparts. We found Mandarin bare nouns occurred as counterparts to both types of German definites, which gave a direct empirical reason to reject Jenks's proposal. The methodology followed in Chapter 7 enabled us to focus on the strong definite contexts in the HP corpus, where the division of labor between bare nouns and demonstratives is unsettled. This precise dataset served as the testing ground for Ahn's (2022) distinction between anaphoric definites and demonstratives. The findings confirm that the distribution of bare nouns and demonstratives in Mandarin was predictable based on factors for anaphoricity, such as referential distance and salience, rather than definiteness itself. These results support the hypothesis that Mandarin demonstratives are canonical demonstratives rather than definiteness markers.

As outlined in Section 1.4, the power of this methodology lies in its design, which anticipates potential criticisms and integrates controls as inherent strengths. We can thus answer the methodological research question:

Methodologically, how can we ensure that a translation corpus approach provides us with reliable data that allow us to answer our empirical questions?

This general question is broken into the following subquestions:

Subquestion 3.1: How can we make sure we do justice to Mandarin expressions *per se* and not only as translations of NPs preceded by *the* and *a*? Subquestion 3.2: How can we make sure that the potential biases of our random corpus selection do not limit the scope of the empirical answers our translation corpus approach can provide? Subquestion 3.3: How can we make sure that translations are sufficiently representative of their target languages? Subquestion 3.4: How can we ensure that meaning differences between originals and translations do not preclude drawing valid conclusions on the basis of translated data?

Chapter 2 employed a general referential system analysis between English and Mandarin rather than pre-selecting the definite and indefinite contexts. This is done to ensure a representation of Mandarin expressions on their own terms, rather than merely as translations of English definite and indefinite expressions (Subquestion

3.1). The inclusion of benchmark categories (e.g., pronouns and possessives) from the whole referential system, whose alignment between English and Mandarin was confirmed to be as expected in the literature, provided independent evidence for the representativeness of the corpus data (Subquestion 3.3). The greatest advance of this methodology is its cross-linguistic extension of the translation corpus. By introducing Russian and Hindi as two other “article-less” languages in Chapter 3, the methodology was able to validate language-specific patterns from potential artifacts of the corpus bias caused by the selection of English source texts (Subquestion 3.2). This multilingual comparison provided additional evidence that the alternation in Mandarin is a genuine feature of its referential system, not a corpus bias. Finally, the translation corpus in Chapter 4 was set up to include languages with well-documented article systems (Spanish, German, and Hebrew). In this way, we were able to confirm the assumption of semantic stability of definiteness and indefiniteness across translation (Subquestion 3.4). These languages behaved exactly as predicted by the literature, thus providing evidence that the core semantics of definiteness and indefiniteness remain robustly stable across translations, which validated the foundational premise of using English articles as semantic proxies for regular definite and indefinite contexts.

In sum, the translation corpus methodology developed in this thesis systematically turns the parallel corpus into a powerful tool for analyzing languages like Mandarin, which lack dedicated markers for (in)definiteness, and for testing theoretical analyses. Thus, the methodology supports the thesis’ key empirical and theoretical contributions on Mandarin encoding of (in)definiteness and the alternation challenge and represents an important advance in the study of referential systems in articleless languages.

8.5 Research limitations and future directions

While this thesis yielded a series of empirical findings and systematic analyses on the basis of translation corpus methodology, it has limitations that also point to directions for future studies. This section addresses three main aspects of these limitations and future corresponding directions for future research.

8.5.1 Potential scope of empirical data

To ensure deep contextual comparability across languages, this study focused on a single, rich parallel corpus, *viz.*, *Harry Potter and the Philosopher’s Stone* and its multilingual translations. While this focus was crucial for the internal validity of our

findings, it also naturally invites future researchers to test the external validity of the observed patterns across a wider range of genres.

The limitation of the approach used in this thesis is that the narrative style of fantasy fiction has specificities in lexical choice and referential patterns. However, as Tellings et al. (2022) have demonstrated using the same *Harry Potter* novel offers a valuable opportunity to distinguish between narrative discourse and fictional dialogue. In their cross-linguistic study on the Present Perfect, Tellings et al. (2022) argued that the dialogue portions of the novel effectively reflect features of spoken language grammar, contrasting sharply with the narrative sections. They showed that grammatical forms can function as indexical categories restricted to dialogue or display significantly different distributions between the two registers. Future research on Mandarin reference can adopt their methodology to separate dialogue from narrative text to investigate register-dependent variations.

Another limitation is related to the study of bare classifier phrases (CI+N), which were not treated as a separate variable in the current study. As noted by Cheng and Sybesma (1999), the distribution of CI+N differs from the *yi*+CI+N construction. While *yi*+CI+N can function as a specific or non-specific indefinite, CI+N typically encodes a strictly indefinite reading. The current thesis does not delve into CI+N since the dataset is very limited, given its potential association with colloquial speech rather than narratives. By applying the fine-grained distinction between narrative and dialogue as established by Tellings et al. (2022), future work can empirically test and analyze the distribution of CI+N in the indefinite contexts in the dialogue portions of the corpus.

Furthermore, future investigations should cover the indefinite domain more comprehensively by incorporating number distinctions. This thesis primarily focused on the alternation between bare nouns and the atomic indefinite marker *yi* ('one'), given that we take the English singular indefinites as the proxies for indefiniteness. However, other indefinite forms, such as *yixie* ('some'), warrant further examination to complete the picture of Mandarin indefiniteness. Recent theoretical developments on the interaction between number and classifiers, such as Doetjes (2021)'s cross-linguistic analysis, provide a promising framework for analyzing these variations. Integrating *yixie* and CI-N within such a framework would provide a deeper understanding of how Mandarin encodes the interplay between indefiniteness and plurality.

8.5.2 Methodological triangulation

The core findings of this thesis are based on a series of translation corpus studies. While our corpus methods have been proved to be effective and reliable for uncovering language patterns, the findings can be further tested through integration with other intra-linguistic corpora or translation-related methods.

For instance, we can conduct psycholinguistic experiments with eye-tracking, self-paced reading, or ERPs to deeply investigate the cognitive load of different patterns with alternation forms (pseudo-incorporated bare nouns *vs.* numeral-*yi* phrases as regular arguments) and with patterns without alterations (bare nouns *vs.* demonstratives) during real-time sentence processing. Such methods could, for instance, provide processing evidence for the proposed distinction between numeral-*yi* phrases as regular arguments and pseudo-incorporated bare nouns, potentially revealing different cognitive loads. We can also conduct quantitative analyses of the distribution frequency of bare nouns, numeral-*yi* constructions, and demonstratives in large-scale native language corpora (e.g., BCC Corpus). This would allow us to verify if the frequency distributions we found in the Harry Potter corpus hold true in a larger, more varied Mandarin corpus. We can also use the key hypotheses proposed in this study—especially as they concern the typicality condition for pseudo-incorporation and the demonstrative as anaphoricity—in translation experiments with minimal pairs or graded acceptability for native speakers to conduct. The integration of these methods would lead to a more robust set of evidence for the empirical/theoretical conclusions of this thesis and for the Mandarin referential system.

8.5.3 Theoretical exploration

The theoretical analysis in this dissertation was conducted within the Properties Approach, which was refined throughout the thesis to account for the complexities of the Mandarin referential system.

The choice of this framework was motivated by cross-linguistic comparisons, most notably those with Russian and Hindi. The adapted Properties Approach offers the flexibility needed to explain the observed cross-linguistic patterns. Since this thesis focuses on how Mandarin encodes definiteness and indefiniteness, the theoretical analysis and proposal testing were primarily restricted to Mandarin. Future work, however, can extend the adapted Properties Approach to Russian and Hindi, particularly in the definite domain, by building on Ahn's (2022) analysis and applying it to these languages.

This theoretical commitment to the adapted Properties Approach does not preclude the possibility that the Kinds Approach could also account for the same phenomena. Future research should systematically assess the potential of the Kinds Approach in light of the empirical findings presented in this thesis. Specifically, it needs to be examined how the Kinds Approach accounts for the alternation between numeral-*yi* and bare nouns in the indefinite domain, as well as the lack of alternation between demonstratives and bare nouns in the definite domain. To capture the cross-linguistic variation, there is also a need to explain the free use of Russian bare singulars in indefinite contexts and the absence of a clear division of labor between bare nouns and demonstratives in the definite domain in both Russian and Hindi. This may require an in-depth reassessment of the hierarchy of type-shifters ($\exists < \iota, \cap$) and the Blocking Principle's operation within the framework.

Further comparative research between the Kinds Approach and the Properties Approach would help us evaluate the theoretical neutrality of the empirical findings and foster a deeper understanding of argument formation across languages by developing competing analyses within distinct theoretical frameworks.

8.6 Conclusion

This thesis began by addressing the alternation challenge regarding how Mandarin expresses definiteness and indefiniteness. Specifically, it addressed the functionalist observation that Mandarin bare nouns alternate with demonstratives in definite contexts and with numeral-*yi* in indefinite contexts, which poses a challenge to formal analyses on Mandarin argument formation. The investigation hinged on a robust methodological foundation. The translation corpus, which used English definite and indefinite articles as semantic proxies for regular definite and indefinite contexts, provided an efficient toolkit for making Mandarin (in)definite contexts analyzable. This methodology licensed a nuanced empirical picture. For the indefinite domain, it confirmed a genuine alternation and revealed a systematic division of labor. Specifically, it showed that numeral-*yi* functions as a regular indefinite article, while bare nouns are restricted to non-argumental, pseudo-incorporated positions. Moreover, it revealed that bare nouns are the default choice in the definite domain, while demonstratives operate in the separate domain of anaphoricity.

These empirical discoveries allowed us to resolve the theoretical puzzle within an adapted Properties Approach. In the indefinite domain, the alternation challenge was resolved by analyzing numeral-*yi* as an indefinite article in argument positions.

According to the pseudo-incorporation analysis, indefinite bare nouns derive their existential force from the verb–noun combination with typicality relation, thus avoiding the blocking effect of numeral-*yi*. In the definite domain, the challenge was resolved by showing that Mandarin demonstratives are not definite articles but canonical anaphoric demonstratives. Therefore, demonstratives no longer block bare nouns from acquiring definiteness through the covert iota type-shift.

Based on these outcomes, this thesis answers the question of how Mandarin encodes definiteness and indefiniteness, thereby reconciling empirical insights of functionalist literature with the explanatory power of formal semantics.

APPENDIX A

Distribution of English and Mandarin Referential
Expressions in the HP Corpus (dataset for Chapter 2)

Table A.1: Distribution of English Referential Expressions from HP1 Chapter 1.

Grammatical Category	Label	Count	Category Count	Category %			
personal pronoun	personal pronoun	445	445	36.8%			
definites	the+N	138	173	14.30%			
	the+N-s	35					
possessives	possessives	117	117	9.67%			
proper name	proper name	86	152	12.56%			
	title	66					
indefinites	a/an+N	90	90	7.44%			
bare noun	N-s	52	89	7.36%			
	N	37					
indefinite pronoun	various (see text)	42	59	4.88%			
	no + N	4					
	any + N	3					
	the other + N	2					
	some + N	2					
	another + N	2					
	such a + N	1					
	such + N	1					
	no + N-s	1					
	any + N-s	1					
	demonstratives	demonstrative pronoun			20	34	2.81%
		demonstrative			14		
several + N-s		2					
most of + N		2					
every + N		2					
all this		2					
a few + N-s		2					
so much of + N		1					
much + N		1					
quantifier		lots of + N-s	1	21	1.74%		
		hundreds of + N-s	1				
		all the + N-s	1				
	all the + N	1					
	all that	1					
	all + N-s	1					
	a lot of + N-s	1					
	a dozen + N-s	1					
	a couple of + N	1					
	special structure	conjunction	14			15	1.24%
		N after N	1				
	numeral	numeral-rest	9			13	1.07%
numeral-ONE		1					
time	enumeration	2	3	0.25%			
	today	1					
	the next few + time expression	1					
	next + time expression	1					
Total			1210				

Table A.2: Distribution of Mandarin Referential Expressions from HP1 Chapter 1.

Grammatical Category	Label	Count	Category Count	Category %
personal pronoun	personal pronoun	313	315	26.03%
	personal pronoun + N	2		
bare noun	N	285	295	24.38%
	N + plural marker men	8		
	N + group marker qún	2		
untranslated	no literal translation	122	216	17.85%
	mismatched translation	94		
proper name	proper name + N	79	163	13.47%
	proper name	79		
	proper name + numeral-yi + classifier	5		
numeral	numeral-yi + classifier + N	53	72	5.95%
	numeral-rest	13		
	numeral-yi + classifier	6		
demonstratives	demonstrative + classifier + N	29	68	5.62%
	demonstrative	17		
	demonstrative + N	16		

Table A.2: – Continued

Grammatical Category	Label	Count	Category Count	Category %
possessives	demonstrative + classifier	6	49	4.05%
	possessive + N	44		
	possessive + demonstrative	3		
	possessive + numeral-yi + classifier + N	1		
	possessive	1		
indefinite expression	indefinite pronoun	6	18	1.49%
	such + N	3		
	such + classifier + N	2		
	some + N	2		
	shenme + N	2		
	another + N	2		
	any + N	1		
	you + N	2		
	conjunction	2		
	N + N	1		
non-direct structure	N + demonstrative + classifier	1	7	0.58%
	classifier + N	1		
	much + N	2		
	yixie + N	1		
	most + N	1		
quantifier			7	0.58%

Table A.2: – Continued

Grammatical Category	Label	Count	Category Count	Category %
	lots + N	1		
	every + N	1		
	all + classifier + N	1		
		Total	1210	

Bibliography

- Aguilar-Guevara, A., Pozas Loyo, J., and Vázquez-Rojas Maldonado, V. (2019). *Definiteness across languages*. Language Science Press.
- Ahn, D. (2022). Indirectly direct: An account of demonstratives and pointing. *Linguistics and Philosophy*, 45(6):1345–1393.
- Alexopoulou, D. and Folli, R. (2011). Indefinite topics and the syntax of nominals in Italian and Greek. In *Proceedings of WCCFL*, volume 28.
- Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106(4):748–765.
- Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of pragmatics*, 16(5):443–463.
- Arkoh, R. and Matthewson, L. (2013). A familiar definite article in Akan. *Lingua*, 123:1–30.
- Barrie, M. and Li, A. (2015). The semantics of (pseudo) incorporation and case. In *The syntax and semantics of pseudo-incorporation*, pages 159–188. Brill.
- Bogaards, M. (2022). The Discovery of Aspect: A heuristic parallel corpus study of ingressive, continuative and resumptive viewpoint aspect. *Languages*, 7(3):158.
- Borik, O. and Gehrke, B. (2015). An introduction to the syntax and semantics of pseudo-incorporation. In *The syntax and semantics of pseudo-incorporation*, pages 1–43. Brill.

- Borik, O., Le Bruyn, B., Liu, J., and Seres, D. (2025). Bare nouns in slavic and beyond. In Gehrke, B., Lenertová, D., Meyer, R., Seres, D., Szucsich, L., and Zaleska, J., editors, *Advances in Formal Slavic Linguistics 2022*, pages 107–128. Language Science Press, Berlin.
- Borthen, K. (2003). *Norwegian bare singulars*. NTNU, Trondheim.
- Bremmers, D., Liu, J., van der Klis, M., and Le Bruyn, B. (2022). Translation Mining: Definiteness across Languages (A Reply to Jenks 2018). *Linguistic Inquiry*, 53(4):735–752.
- Bronnikov, G. (2006). A critique of Dayal (2004). Term Paper, University of Texas at Austin.
- Carlson, G. N. (1977). A unified analysis of the English bare plural. *Linguistics and Philosophy*, 1(3):413–457.
- Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press, Berkeley, CA.
- Chen, P. (2003). Indefinite determiner introducing definite referent: a special use of ‘yi ‘one’+ classifier’ in chinese. *Lingua*, 113(12):1169–1184.
- Chen, P. (2004). Identifiability and definiteness in chinese. *Linguistics*, 42(6):1129–1184.
- Cheng, L. L.-S. and Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry*, 30(4):509–542.
- Chierchia, G. (1998). Reference to kinds across language. *Natural Language Semantics*, 6(4):339–405.
- Chierchia, G. (2010). Mass nouns, vagueness and semantic variation. *Synthese*, 174:99–149.
- Coppock, E. and Beaver, D. (2014). A superlative argument for a minimal theory of definiteness. In *Semantics and Linguistic Theory*, pages 177–196.
- Dayal, V. (2003). A semantics for pseudo-incorporation. *Ms., Rutgers University*.
- Dayal, V. (2004). Number marking and (in) definiteness in kind terms. *Linguistics and Philosophy*, 27:393–450.

- Dayal, V. (2011). Hindi pseudo-incorporation. *Natural Language & Linguistic Theory*, 29(1):123–167.
- Dayal, V. (2015). Incorporation: Morpho-syntactic vs. semantic considerations. In *The syntax and semantics of pseudo-incorporation*, pages 47–87. Brill.
- Dayal, V. and Jiang, L. J. (2022). The puzzle of anaphoric bare nouns in mandarin: A counterpoint to index! *Linguistic inquiry*, 54(1):147–167.
- de Swart, H. (2020). Double negation readings. In *The Oxford Handbook of Negation*. Oxford University Press.
- de Swart, H., Grisot, C., Le Bruyn, B., and Xiqués, T. M. (2022a). Perfect variations in Romance. *Isogloss: Open Journal of Romance Linguistics*, 8(5):1–31.
- de Swart, H., Tellings, J., and Wälchli, B. (2022b). Not... until across european languages: A parallel corpus study. *Languages*, 7(1):56.
- Diessel, H. (1999). *Demonstratives*. John Benjamins Publishing Company.
- Dobrovie-Sorin, C., Bleam, T., and Espinal, M. T. (2006). Bare nouns, number and types of incorporation. In Vogeleer, S. and Tasmowski, L., editors, *Non-definiteness and plurality*, pages 51–79. John Benjamins, Amsterdam.
- Doetjes, J. S. (2021). Number and numeral classifiers. In Cabredo Hofherr, P., editor, *The Oxford Handbook of Grammatical Number*, pages 220–241. Oxford University Press, Oxford.
- Doron, E. (2003). Bare singular reference to kinds. *Semantics and Linguistic Theory*, 13:73–90.
- Dowty, D. R. (1979). *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*, volume 7. Springer Science & Business Media.
- Ebert, K. (1971a). *Referenz, Sprechsituation und die bestimmten Artikel in einem nordfriesischen Dialekt (Fering)*. PhD thesis, Christian-Albrechts-Universität zu Kiel, Kiel.
- Ebert, K. (1971b). Zwei Formen des bestimmten Artikels. In Wunderlich, D., editor, *Probleme und Fortschritte der Transformationsgrammatik*, pages 159–174. Hueber, Munich.

- Elbourne, P. (2008). Demonstratives as individual concepts. *Linguistics and philosophy*, 31(4):409–466.
- Epstein, R. (1993). The definite article: early stages of development. In van Marle, J., editor, *Historical Linguistics 1991: Papers from the 10th International Conference on Historical Linguistics, Amsterdam, August 12–16, 1991*, pages 111–134. John Benjamins, Amsterdam/Philadelphia.
- Espinal, M. T. (2010). Bare nominals in Catalan and Spanish. their structure and meaning. *Lingua*, 120(4):984–1009.
- Espinal, M. T. and McNally, L. (2011). Bare nominals and incorporating verbs in Spanish and Catalan. *Journal of Linguistics*, 47(1):87–128.
- Farkas, D. F. and Swart, H. d. (2003). *The semantics of incorporation: From argument structure to discourse transparency*. University of Chicago Press.
- Fuchs, M. and González, P. (2022). Perfect-perfective variation across spanish dialects: a parallel-corpus study. *Languages*, 7(3):166.
- Gehrke, B. (2022). Differences between Russian and Czech in the use of aspect in narrative discourse and factual contexts. *Languages*, 7(2):155.
- Gehrke, B. and Lekakou, M. (2013). How to miss your preposition. *Studies in Greek linguistics*, 33:92–106.
- Georgakopoulos, T. and Polis, S. (2018). The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, 12(2):e12270.
- Givón, T. (1981). On the development of the numeral ‘one’ as an indefinite marker. *Folia linguistica historica*, 15(Historica-vol-2-1):35–54.
- Givón, T. (1984/1990). *Syntax: A functional-typological introduction*. John Benjamins.
- Greenberg, J. H. (1978). How does a language acquire gender markers? In Greenberg, J. H., Ferguson, C. A., and Moravcsik, E. A., editors, *Universals of Human Language, Volume 3: Word Structure*, pages 47–82. Stanford University Press, Stanford, CA.

- Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.
- Hawkins, J. A. (1978). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Croom Helm, London.
- Heim, I. (1991). Artikel und Definitheit. In von Stechow, A. and Wunderlich, D., editors, *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 487–535. Walter de Gruyter, Berlin.
- Hsiao, C.-h. (2011). Personal pronoun interchanges in mandarin chinese conversation. *Language Sciences*, 33(5):799–821.
- Huang, C.-T. J. (2015). On syntactic analyticity and parametric theory. In Li, A. Y.-H., Simpson, A., and Tsai, W.-T., editors, *Chinese syntax in a cross-linguistic perspective*, pages 1–48. Oxford University Press, Oxford.
- Huang, S. (1999). The emergence of a grammatical category definite article in spoken chinese. *Journal of pragmatics*, 31(1):77–94.
- Jenks, P. (2018). Articulated definiteness without articles. *Linguistic Inquiry*, 49(3):501–536.
- Jiang, L. and Dayal, V. (2023). The puzzle of anaphoric bare nouns in Mandarin: A counterpoint to Index! *Linguistic Inquiry*, 54(1):147–167.
- Jiang, L. J. (2017). Mandarin associative plural-men and NPs with-men. *International Journal of Chinese Linguistics*, 4(2):191–256.
- Jiang, L. J. (2020). *Nominal arguments and language variation*. Oxford University Press, Oxford.
- Kaiser, E. (2003). *The quest for a referent: A crosslinguistic look at reference resolution*. PhD thesis, University of Pennsylvania, Philadelphia, PA.
- Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes from Kaplan*, pages 481–563. Oxford University Press, Oxford.

- Kiefer, F. (1990). Noun incorporation in hungarian. *Acta Linguistica Hungarica*, 40(1/2):149–177.
- Kiefer, F. and Németh, B. (2019). Compounds and multi-word expressions in hungarian. *Complex Lexical Units*, page 337.
- King, J. C. (2001). *Complex demonstratives: A quantificational account*, volume 2. Mit Press.
- Kratzer, A. (1996). Severing the external argument from its verb. In Rooryck, J. and Zaring, L., editors, *Phrase structure and the lexicon*, pages 109–137. Springer, Dordrecht.
- Krifka, M. (2003). Bare NPs: Kind-referring, indefinites, both, or neither? In *Proceedings of Semantics and Linguistic Theory 13*, pages 180–203.
- Lazaridou-Chatzigoga, D. (2011). The distribution and interpretation of bare singular count nouns in greek. In *Workshop on Weak Referentiality*.
- Le Bruyn, B. and de Swart, H. (2022). Exceptional wide scope of bare nominals. *Semantics and Pragmatics*, 15:7.
- Le Bruyn, B. and de Swart, H. (2023). Introduction: Tense and aspect across languages. *Languages*, 8(1):33.
- Le Bruyn, B. and de Swart, H. (2024). Cross-linguistic research, parallel corpora, and replication in the Translation Mining Tradition. In *Crosslinguistic Approaches to Language Analysis*, pages 27–55. Cambridge Scholars Publishing.
- Le Bruyn, B., de Swart, H., and Zwarts, J. (2016). From HAVE to HAVE-verbs: Relations and incorporation. *Lingua*, 182:49–68.
- Le Bruyn, B., Fuchs, M., van der Klis, M., Liu, J., Mo, C., Tellings, J., and De Swart, H. (2022). Parallel corpus research and target language representativeness: The contrastive, typological, and translation mining traditions. *Languages*, 7(3):176.
- Le Bruyn, B., van der Klis, M., and de Swart, H. (2024). The HAVE-PERFECT and the tense-aspect grammar of western European languages. *Beyond Aspectual Semantics: Explorations in the Pragmatic and Cognitive Realms of Aspect*, page 143.

- Li, C. N. and Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. University of California Press, Berkeley, CA.
- Liu, J., Dong, X., and Le Bruyn, B. (2022). Mandarin bare indefinites. In Bary, C. and Maier, M., editors, *Proceedings of Sinn und Bedeutung 26*, pages 575–591.
- Liu, J., Patil, S., Schurr, H., Seres, D., Borik, O., and Le Bruyn, B. (2023a). The theory of argument formation: between kinds and properties. In *Poster presented at Semantics and Linguistic Theory 33*.
- Liu, J., Patil, S., Seres, D., Borik, O., and Le Bruyn, B. (2023b). 'Articleless' languages are not created equal. In *Proceedings of Sinn und Bedeutung 27*, pages 381–398, Prague. Charles University.
- Löbner, S. (2011). Concept types and determination. *Journal of Semantics*, 28(3):279–333.
- Longobardi, G. (1994). Reference and proper names: A theory of n-movement in syntax and logical form. *Linguistic Inquiry*, 25(4):609–665.
- Lü, S. (1947). *Essentials of Chinese grammar*. Commercial Press.
- Lü, S. (1955). *Hanyu yufa lunwenji (Collected essays on Chinese grammar)*. Beijing: Kexue Chubanshe.
- Lü, S. (1979). *Hanyu yufa fenxi wenti [Issues in the analysis of Chinese grammar]*. Shangwu Yinshuguan, Beijing.
- Luo, Q. (2022). Bare nouns, incorporation, and event kinds in Mandarin Chinese. *Journal of East Asian Linguistics*, 31(2):221–263.
- Massam, D. (2001). Pseudo noun incorporation in Niuean. *Natural Language & Linguistic Theory*, 19(1):153–197.
- Massam, D. (2020). *Niuean: Predicates and arguments in an isolating language*, volume 6. Oxford University Press.
- McKenzie, A. R. (2012). *The role of contextual restriction in reference-tracking*. University of Massachusetts Amherst.
- McKenzie, A. R. (2015). Deriving topic effects in Kiowa with semantics and pragmatics. *Methodologies in semantic fieldwork*, page 269.

- Meier, C. (2019). Temporal information and definite descriptions. In Blake, K., Davis, F., Lamp, K., and Rhyne, J., editors, *Proceedings of SALT 29*, Los Angeles, CA. Linguistic Society of America. Extended abstract/poster.
- Mo, C. (2022). *The Compositionality of Mandarin Aspect: A Parallel Corpus Study*. PhD thesis, Utrecht University, Utrecht.
- Mueller-Reichau, O. (2015). Pseudo-incorporation in Russian? Aspectual competition and bare singular interpretation. In Borik, O. and Gehrke, B., editors, *The Syntax and Semantics of Pseudo-Incorporation*, pages 262–295. Brill, Leiden.
- Mulder, G., Schoenmakers, G.-J., Hoenselaar, O., and de Hoop, H. (2022). Tense and aspect in a Spanish literary work and its translations. *Languages*, 7(3):217.
- Niu, F. (2015). *Nominal Possession in Mandarin Chinese*. PhD thesis, Queen Mary University of London.
- Partee, B. H. (1987). Noun phrase interpretation and type-shifting principles. In Groenendijk, J., de Jongh, D., and Stockhof, M., editors, *Studies in discourse representation theory and the theory of generalized quantifiers*, pages 115–143. Foris, Dordrecht.
- Partee, B. H. (1999). Weak NP's in HAVE sentences. *JFAK, a Liber Amicorum for Johan van Benthem on the occasion of his 50th Birthday*, pages 39–57.
- Partee, B. H. (2008). A note on mandarin possessives, demonstratives, and definiteness. In *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, pages 263–280. John Benjamins Publishing Company.
- Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.
- Saha, A., Saĝ, Y., Cui, J., and Davidson, K. (2024). Anaphoric demonstratives in mandarin. In *Semantics and Linguistic Theory*, pages 213–232.
- Schwarz, F. (2009). *Two types of definites in natural language*. PhD thesis, University of Massachusetts Amherst, Amherst, MA.
- Schwarz, F. (2014). How weak and how definite are weak definites? In *Weak referentiality*, pages 213–235. John Benjamins Publishing Company.

- Seres, D. and Borik, O. (2018). Definiteness in the absence of uniqueness: The case of Russian. *Advances in formal Slavic linguistics*, pages 339–363.
- Šimík, R. and Demian, C. (2020). Definiteness, uniqueness, and maximality in languages with and without articles. *Journal of Semantics*, 37(3):311–366.
- Simpson, A. and Wu, Z. (2022). Constraints on the representation of anaphoric definiteness in mandarin chinese: A reassessment. In *New explorations in Chinese theoretical syntax*, pages 301–330. John Benjamins Publishing Company.
- Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of york english: Was/were variation as a case study for statistical practice. *Language variation and change*, 24(2):135–178.
- Tellings, J. and Fuchs, M. (2021). Sluicing and temporal definiteness. *Manuscript. Utrecht University*.
- Tellings, J., Fuchs, M., van der Klis, M., Le Bruyn, B., and De Swart, H. (2022). Perfect variations in dialogue: a parallel corpus approach. In *Semantics and Linguistic Theory*, pages 22–43.
- van der Klis, M., Le Bruyn, B., and de Swart, H. (2022). A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse. *Journal of Linguistics*, 58(2):423–457.
- van der Klis, M., Le Bruyn, B., and de Swart, H. E. (2020). De la sémantique des temps verbaux à la traductologie: une comparaison multilingue de l'étranger de camus. In *The Expression of Tense, Aspect, Modality and Evidentiality in Albert Camus's L'Étranger and Its Translations/L'Étranger de Camus et ses traductions: questions de temps, d'aspect, de modalité et d'évidentialité (TAME)*, pages 11–38. John Benjamins Publishing Company.
- Vikner, C. and Jensen, P. A. (2002). A semantic analysis of the English genitive: Interaction of lexical and compositional semantics. *Studia Linguistica*, 56(2):191–226.
- Wälchli, B. and Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, (3):671–710.
- Wiltschko, M. (2013). Descriptive relative clauses in Austro-Bavarian German. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 58(2):157–189.

- Wolter, L. (2006). *That's That: The semantics and pragmatics of demonstrative noun phrases*. PhD thesis, University of California, Santa Cruz, Santa Cruz, CA.
- Wright, S. E. and Givón, T. (1987). The pragmatics of indefinite reference: Quantified text-based studies. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 11(1):1–33.
- Xiang, X. (2019). Personal pronouns in chinese discourse. In *The Routledge handbook of Chinese discourse analysis*, pages 147–159. Routledge.
- Zhang, N. N. (2015). Nominal-internal phrasal movement in mandarin chinese. *The Linguistic Review*, 32(2):375–425.
- Zhu, D. (1982). *Yufa jiangyi [Lectures on grammar]*. Shangwu Yinshuguan, Beijing.

Source Text and Translations

Rowling, J. K. (1997). *Harry Potter and the philosopher's stone*. Bloomsbury, London.

Rowling, J. K. (1997/1998). *Harry Potter und der Stein der Weisen*. Carlsen Verlag, Hamburg. Translated by Klaus Fritz.

Rowling, J. K. (1999). *Hari Poter ve-Even Ha-hakhamim*. Yediot Ahronoth, Rishon LeZion. Translated by Gili Bar-Hillel.

Rowling, J. K. (2000). *Hā lì bō tè yǔ mó fǎ shí*. People's Literature Publishing House, Beijing. Translated by Su Nong.

Rowling, J. K. (2001). *Garri Potter i filosofskiy kamen*. Rosmèn-Press, Moskva. Translated by Igor Oranskij.

Rowling, J. K. (2003). *Hairī Poṭar aur Pāras Patthar*. Manjul Publishing House, Bhopal. Translated by Dr. Sudhir Dixit.

Samenvatting in het Nederlands

Een fundamentele vraag binnen de linguïstiek is hoe talen referentie coderen, specifiek het onderscheid tussen bepaaldheid en onbepaaldheid. Terwijl talen zoals het Engels lidwoorden (*the, alan*) gebruiken om dit onderscheid te markeren, zijn er talen zoals het Mandarijn-Chinees die lijken te functioneren zonder een specifiek lidwoorden-systeem. Dit vormt een puzzel voor de semantiek: hoe drukt het Mandarijn-Chinees bepaaldheid en onbepaaldheid uit bij afwezigheid van expliciete morfologische markeringsen?

Dit proefschrift onderzoekt dit raadsel middels een grondige bestudering van het Mandarijn-Chinees (hierna: Mandarijn). De centrale spanning die dit onderzoek drijft, is de discrepantie tussen twee dominante tradities in de linguïstiek: de functionalistische traditie en de formeel-semantische traditie. De functionalistische literatuur heeft lang geobserveerd dat, hoewel zelfstandige naamwoorden in het Mandarijn 'kaal' (bare) lijken te zijn (zonder determinatoren voor het naamwoord), de taal ook specifieke strategieën hanteert die lijken op het gebruik van lidwoorden in andere talen. Specifiek merken functionalisten op dat kale naamwoorden alterneren met de 'telwoord-één (*yi*)'-constructie (*yi* + classifier + naamwoord) in onbepaalde contexten, en met de constructie met aanwijzende voornaamwoorden (*zhe/na*)-constructie (*zhe/na* + classifier + naamwoord) in bepaalde contexten. Functionalisten betogen dat deze markeerders in deze contexten een proces van semantische verbleking (semantic bleaching) ondergaan, waarbij ze hun canonieke kardinale of deictische kracht verliezen om respectievelijk te fungeren als markeerders van onbepaaldheid en bepaaldheid. Omgekeerd stelt de formele literatuur over argumentsvorming dat kale naamwoorden in het Mandarijn geen extra formele ondersteuning nodig hebben om als argumenten te fungeren. De twee dominante theoretische kaders, de Kinds Approach (Chierchia 1998)

en de Properties Approach (Krifka 2004), analyseren kale naamwoorden in het Mandarijn als verwijzend naar soorten (kinds, type <e>) of eigenschappen (properties, type <e,t>) die covert typeverschuiving (covert type-shifting) ondergaan om als bepaalde of onbepaalde argumenten te fungeren. Bijgevolg voorspellen deze kaders dat het telwoord-*yi* en aanwijzende determinatoren geen rol spelen in standaard bepaalde en onbepaalde contexten, en dat hun voorkomen enkel verwacht wordt in canonieke contexten voor telwoorden of aanwijzende determinatoren, zoals die waarin kardinaliteit of deixis van belang zijn.

De functionalistische empirische claim dat kale naamwoorden alterneren met de ‘telwoord-één’-constructie en de constructie met aanwijzende determinatoren, wordt in dit proefschrift geïdentificeerd als de ‘alternantie-uitdaging’ (alternation challenge) voor de formeel-semantiche traditie. Om deze uitdaging op te lossen, maakt dit onderzoek gebruik van een datagestuurde vertaalcorpusmethodologie. De kern wordt gevormd door J.K. Rowlings *Harry Potter and the Philosopher’s Stone* en de Mandarijnse vertaling daarvan. De studie gebruikt vertalingen van NP’s die worden ingeleid door *the* en *alan* als semantische graadmeters voor ‘reguliere’ bepaalde en onbepaalde contexten. Dit maakt een reeks kwantitatieve analyses van de distributie van Mandarijnse referentiële vormen in deze contexten mogelijk, om de geldigheid van functionalistische claims en formalistische voorspellingen te toetsen.

Hoofdstuk 2 legt eerst de empirische basis door de daadwerkelijke distributie van de betreffende Mandarijnse vormen in bepaalde en onbepaalde contexten te onderzoeken. We valideren eerst de betrouwbaarheid van het corpus door benchmarkcategorieën (voornaamwoorden en bezittelijke vormen) te analyseren waarvan de correspondentie tussen Engels en Mandarijn in de literatuur vaststaat. Vervolgens lijnen we het gehele referentiële systeem tussen Engels en Mandarijn uit ($n = 1210$). Deze brede reikwijdte was cruciaal om de bias van voorselectie, gebaseerd op enkel Engelse lidwoorden, te vermijden en zo te verzekeren dat het geobserveerde patroon het Mandarijnse referentiële systeem op zichzelf weerspiegelt. Wat onze kerndata betreft, tonen de resultaten aan dat er hybride patronen voorkomen in bepaalde en onbepaalde contexten in het Mandarijn. In het bepaalde domein vinden we dat zowel kale naamwoorden als aanwijzende determinatoren voorkomen: Mandarijnse kale naamwoorden zijn de meerderheidskeuze (81%) en Mandarijnse aanwijzende determinatoren verschijnen in 14% van de gevallen. In het onbepaalde domein vinden we ook dat zowel kale naamwoorden als het telwoord-*yi* voorkomen: het Mandarijnse telwoord-*yi* is de dominante keuze (65%) en kale naamwoorden verschijnen in slechts 26% van de gevallen en zijn grotendeels beperkt tot objectposities. Deze resultaten beves-

tigen de functionalistische observatie van hybride systemen. Voordat we deze puzzels empirisch en theoretisch verkennen, is onze volgende stap te bevestigen dat het geobserveerde Mandarijnse patroon geen artefact is van de Engelse bronteksten, maar een authentiek Mandarijn-specifiek patroon.

In hoofdstuk 3 breiden we, om te onderzoeken of de onderscheiden tussen Mandarijnse kale naamwoorden en andere vormen zoals geobserveerd in hoofdstuk 2 iets indiceren, het kerncorpus uit met andere lidwoordloze talen (*article-less languages*), namelijk Russisch en Hindi naast het Mandarijn, in hun vertalingen van hetzelfde Harry Potter-corpus als toegepast in hoofdstuk 2. Als de resultaten van hoofdstuk 2 betreffende het Mandarijnse patroon ad-hoc resultaten zouden zijn van de Engelse bronteksten, zouden we in het Russisch en Hindi hetzelfde patroon moeten vinden als in het Mandarijn, aangezien deze allen geclassificeerd zijn als ‘lidwoordloos’ (Dayal 2004). De vergelijking toont aan dat deze ‘lidwoordloze’ talen een duidelijke divergentie laten zien in hoe ze omgaan met onbepaalde contexten. Terwijl het Russisch bijna uitsluitend vertrouwt op kale naamwoorden (meer dan 80% van de gevallen) en het telwoord-één in slechts 5% van de gevallen gebruikt, gebruikt het Mandarijn het telwoord-*yi* in 66% van dezelfde contexten. Het Hindi neemt een middenpositie in en gebruikt het telwoord-één in 40% van de onbepaalde contexten. Een Fisher’s Exact Test bevestigde dat het verschil tussen het Mandarijn en het Russisch statistisch significant is. In bepaalde contexten zijn de resultaten echter uniformer. Hoewel alle drie de talen voornamelijk vertrouwen op kale naamwoorden voor het Engelse *the+N*, toonde de statistische analyse een significant verschil aan tussen het Mandarijn en het Russisch, maar niet tussen de andere taalparen. Deze taaloverstijgende corpus-resultaten stellen vast dat in het onbepaalde domein de Mandarijnse dominantie van het telwoord-*yi* en de minderheid van kale naamwoorden een taalspecifiek kenmerk is, terwijl in het bepaalde domein het onderscheid tussen Mandarijnse aanwijzende determinatoren en kale naamwoorden een subtielere distributionele tendens blijft.

Hoofdstuk 4 evalueert de verklarende kracht van de twee dominante formele kaders: de Kinds Approach (KA) en de Properties Approach (PA), om de theoretische basis te leggen voor een formeel-semantische oplossing voor de ‘alternantie-uitdaging’ (*alternation challenge*). Beide kaders worden geconfronteerd met theoretische uitdagingen bij het verklaren van het geobserveerde Mandarijnse patroon. Specifiek stelt de PA dat een openlijke markeerder zou moeten blokkeren (*block*) dat een kaal naamwoord via een coverte typeverschuiving dezelfde bepaalde of onbepaalde betekenissen verkrijgt. Omgekeerd kampt de KA met een gebrek aan motivatie: als Mandarijnse kale naamwoorden van nature argumenteel zijn, is er geen theoretische druk voor de

taal om lidwoordachtige openlijke uitdrukkingen te ontwikkelen. Om te verzekeren dat de theoretische evaluatie gegrond is in een uitgebreide typologie, werpen we het empirische net nog wijder uit door het Harry Potter-corpus uit te breiden met vertalingen in talen met volledige lidwoordensystemen (Spaans, Duits) en een systeem met enkel een bepaald lidwoord (Hebreeuws). Door het uitlijningspatroon tussen deze talen in hun vertalingen van het Engelse *the* en *alan* te vergelijken, merken we op dat de standaardversies van beide kaders moeite hebben om de taaloverstijgende variaties te verklaren. Tegen het einde van hoofdstuk 4 stelt onze oplossing een aanpassing van de Properties Approach in dit proefschrift voor, waarbij we betogen dat het verwachte blokkeringseffect wordt omzeild doordat kale naamwoorden en openlijke markeerders verschillende semantische of syntactische niches bezetten. Wat betreft onbepaalde contexten in het Mandarijn, rust ons voorstel op twee belangrijke hypothesen: ten eerste, dat het Mandarijnse telwoord-*yi* een onbepaald lidwoord is; en ten tweede, dat kale naamwoorden in onbepaalde contexten pseudo-geïncorporeerd (pseudo-incorporated) zijn. Wat betreft bepaalde contexten in het Mandarijn, toetsen we eerst het voorstel van Jenks (2018) over de taakverdeling tussen Mandarijnse kale naamwoorden en aanwijzende determinatoren als zwakke versus sterke definiten (*weak vs strong definites*). Vervolgens stellen we een nieuwe taakverdeling voor tussen Mandarijnse kale naamwoorden en aanwijzende determinatoren, gebaseerd op Ahns (2022) analyse van anaforische definiten en anaforische demonstrativa.

Hoofdstuk 5 toetst de hypothese dat Mandarijnse onbepaalde kale naamwoorden pseudo-geïncorporeerd (pseudo-incorporated) zijn. Hoewel het Mandarijn expliciete morfosyntactische markeerders voor dit proces ontbeert, overbruggen we de kloof door een nieuwe analyse van Mandarijnse pseudo-incorporatie te ontwikkelen, gebaseerd op de typicaliteit (*typicality*) in werkwoord-naamwoordcombinaties. We volgen de intuïtie van Huang (2015) en stellen typicaliteit in de werkwoord-naamwoordcombinaties voor als het definiërende criterium voor Mandarijnse pseudo-incorporatie. Deze hypothese stelt dat een kaal naamwoord alleen pseudo-geïncorporeerd kan worden als het in een stereotiepe relatie (*stereotypical relation*) staat tot het werkwoord waarmee het combineert; een dergelijke typicaliteitsrelatie komt voort uit de QUALIA-rollen (*QUALIA roles*) van het naamwoord (Pustejovsky 1996). Bijvoorbeeld: ‘boek’ kan voorkomen met ‘lezen’ in een pseudo-incorporatieconstructie omdat de telische rol (*telic role*) van ‘boek’ kan worden opgepikt door het werkwoord ‘lezen’. Ter vergelijking: ‘boek’ kan niet voorkomen in een pseudo-incorporatieconstructie met ‘trekken (van onder een kussen)’, omdat er geen QUALIA-rol van ‘boek’ is die ‘trekken’ kan oppikken. ‘Trek boek (van onder een kussen)’ moet daarom een reg-

uliere existentiële typeverschuiving (*existential type-shift*) ondergaan, wat leidt tot de insertie van het telwoord-*yi*, dat wij beschouwen als het Mandarijnse onbepaalde lidwoord. We voeren een gerichte corpusstudie uit naar alle Engelse *a/an* in de objectposities die vertaald zijn naar het Mandarijn in het Harry Potter-corpus ($n = 154$). We annoteren op typicaliteit van de werkwoord-naamwoordcombinatie in de Mandarijnse data. De resultaten bevestigen dat typicaliteit de primaire factor is voor de taakverdeling tussen kale naamwoorden en het telwoord-*yi* in reguliere onbepaalde contexten. Mandarijnse kale naamwoorden zijn dominant wanneer er sprake is van typicaliteit in de werkwoord-naamwoordcombinatie; elders is het telwoord-*yi* dominant en obligatoir. Dit lost de onbepaaldheidspuzzel op met de empirische validatie van onze hypothese: het telwoord-*yi* fungeert als het onbepaalde lidwoord voor reguliere argumentposities, terwijl kale naamwoorden beperkt zijn tot pseudo-incorporatieconstructies die geregeerd worden door de typicaliteitsrelatie binnen de werkwoord-naamwoord-combinatie.

In hoofdstuk 6 gaan we over naar het bepaalde domein. Zoals uiteengezet in hoofdstuk 4, behandelen we eerst het voorstel van Jenks (2018) in hoofdstuk 6 en verkennen we vervolgens het voorstel van Ahn (2022) in hoofdstuk 7. In hoofdstuk 6 toetsen we het voorstel van Jenks (2018), volgens hetwelk de taakverdeling tussen Mandarijnse kale naamwoorden en aanwijzende determinatoren het semantische onderscheid weerspiegelt tussen zwakke bepaaldheid (*weak definiteness*, uniciteit) en sterke bepaaldheid (*strong definiteness*, anaforiciteit), vergelijkbaar met het Duitse onderscheid tussen zwakke definiëten (als samengetrokken lidwoorden in prepositionele frases) versus sterke definiëten (als niet-samengetrokken lidwoorden in prepositionele frases), in navolging van Schwarz (2009). We voeren een gerichte corpusstudie uit gebaseerd op de Harry Potter-roman, waarbij we Duitse vormen als samengetrokken (zwakke) definiëten en niet-samengetrokken (sterke) definiëten uitlijnen met Mandarijnse vormen van kale naamwoorden en aanwijzende determinatoren. Onze resultaten falsificeren Jenks' hypothese empirisch. Duitse zwakke definiëten lijnen zoals verwacht uit met Mandarijnse kale naamwoorden. Echter, Duitse sterke definiëten lijnen in 18% van de gevallen uit met Mandarijnse aanwijzende determinatoren en in 80% van de gevallen met Mandarijnse kale naamwoorden. Dit indiceert dat Mandarijnse kale naamwoorden niet beperkt zijn tot zwakke bepaaldheid, in tegenstelling tot de voorspellingen van de analyse van het tweeledige lidwoordensysteem.

Nu Jenks' hypothese weerlegd is, toetsen we in hoofdstuk 7 de alternatieve nulhypothese: dat er geen werkelijke alternantie bestaat in de zin van competitie in de reguliere bepaalde contexten tussen kale naamwoorden en aanwijzende determina-

toren. We passen Ahn's (2022) kader voor de analyse van anaforische defnieten en anaforische demonstrativa aan, en stellen dat Mandarijnse aanwijzende determinatoren fungeren als canonieke anaforische demonstrativa die anaforiciteitasserteren, terwijl Mandarijnse kale naamwoorden fungeren als anaforische defnieten die anaforiciteitpresupponeren. We voeren een gerichte corpusstudie uit waarbij we sterke (anaforische) bepaalde contexten in het Mandarijn analyseren ($n = 64$), en annoteren op factoren die aanwijzende determinatoren favoriseren. Onze resultaten tonen aan dat kale naamwoorden het standaardmechanisme blijven voor het coderen van bepaaldheid. Aanwijzende determinatoren worden alleen toegepast wanneer bepaalde pragmatische factoren in het spel komen om anaforiciteit te adresseren. Dit lost derhalve de bepaaldheidspuzzel op: aanwijzende determinatoren concurreren niet in het bepaalde domein, maar behouden hun canonieke rol als markeerders van geasserteerde anaforiciteit (*asserted anaphoricity*), terwijl kale naamwoorden het standaardmechanisme blijven voor het coderen van (sterke) bepaaldheid.

Hoofdstuk 8 besluit dit proefschrift. We betogen dat het Mandarijn in het onbepaalde domein een openlijk lidwoord (*overt article*) heeft ontwikkeld, namelijk het telwoord-*yi*, dat reguliere kale naamwoorden blokkeert en hen beperkt tot pseudo-incorporatieconstructies. In het bepaalde domein blijven kale naamwoorden de standaardkeuze voor bepaalde uitdrukkingen, terwijl aanwijzende determinatoren hun canonieke demonstratieve semantiek behouden. Door deze empirische bevindingen te integreren in een aangepaste Properties Approach, biedt dit proefschrift een robuust theoretisch model dat de complexe realiteit van het Mandarijnse referentiële systeem verklaart en de vertaalcorpusmethodologie valideert als een krachtig instrument voor semantisch onderzoek.

Curriculum Vitae

Jianan Liu was born in Lanzhou, China. She developed a strong interest in linguistics at Lanzhou University before moving to the Netherlands to deepen her theoretical foundations. After earning a Research Master's degree from Utrecht University in 2018, she has been a PhD candidate at the Utrecht Institute of Linguistics OTS (UiL OTS) since 2019. Her doctoral research investigates Mandarin referential systems, focusing on expressions of (in)definiteness through cross-linguistic parallel corpora analysis. This dissertation presents the results of this research.