How radical is pro-drop in Mandarin?

A quantitative corpus study on referential choice in Mandarin Chinese

Masterarbeit
im Masterstudiengang
General Linguistics
in der Fakultät Geistes- und
Kulturwissenschaften
der Otto-Friedrich-Universität
Bamberg

Verfasserin: Maria Carina Vollmer
Prüfer: Prof. Dr. Geoffrey Haig
Zweitprüferin: PD Dr. Sonja Zeman

Contents

List of Abbreviations

1SG	First person singular
2SG	Second person singular
3SG	Third person singular
1PL	First person plural
2PL	Second person plural
3PL	Third person plural
ADP	Adposition
ADV	Adverb
ASP	Aspectual marker
CL	Classifier
DEM	Demonstrative
INCL	Inclusive
MOD	Modifier
MP	Modal particle
NEG	Marker of negation
NC	Non-classifiable
NP	Noun phrase
REFL	Reflexive

List of Figures

1 | Introduction

One of the most famous characteristics of Mandarin Chinese is its characterisation as a radical pro-drop language (Roberts & Holmberg 2009: 9, Neeleman & Szendröi 2007, Liu 2014). The term PRO-DROP refers to the phenomenon when languages admit the possibility that referential forms in in a clause are not realised overtly but are dropped instead. The hearer thus has to infer from context to which referent the argument refers. While other languages also exhibit this so-called PRO-DROP, Mandarin supposedly has a much wider scope of zero arguments and makes extensive use of them. This phenomenon is often referred to as RADICAL PRO-DROP (Roberts & Holmberg 2009: 8ff, Ackema & Neeleman 2007: 83 Ackema et al. 2006: 5, Barbosa 2011a, Neeleman & Szendröi 2007, Liu 2014).

Studies of PRO-DROP are important in many respects. First of all, PRO-DROP is used as a characteristic in typological classifications of languages in several regards, i.e. in the context of TOPIC-PROMINENT languages (Li & Thompson 1976) and NON-CONFIGURATIONALITY (Hale 1983: 5). Ackema et al. (2006: 16) comment that

> [...] one may even wonder whether pro-drop languages do in fact have a structural subject position, or whether in such languages apparent subjects are really optional additions to the clause in a dislocated position. If so, this would be

> reminiscent of the behaviour of all syntactic noun phrases in non-configurational languages. The question is, then, to what extent pro-drop languages are non-configurational.

Understanding pro-drop and referential choice also plays a role in understanding how to identify the referent of an omitted or pronominal argument for machine translation or research on the processing costs of anaphoric expressions (e.g. Gelormini-Lezama 2018). It is thus also relevant in computer sciences, since it is crucial for translation machines to correctly identify the referent of a pronoun or dropped argument (see e.g. Zhang et al. 2019, Wang et al. 2017, Soares 2016 for research on machine translation between pro-drop and non-pro-drop languages).

Even though pro-drop is relevant in many respects, there has been no large-scale quantitative study on the frequency of zero arguments in Mandarin in comparison to a larger group of typologically and geographically diverse languages. This kind of study has only become possible in recent years, since larger corpora of languages with consistent annotations (of zero arguments) have only become available now. This thesis responds to this research gap and the new possibilities with regard to corpus studies and statistical software, and aims at quantitatively analysing the frequency of zero arguments in Mandarin Chinese in natural spoken language in comparison to other languages, using Multi-CAST, the *The Multilingual Corpus of Annotated Spoken Texts* (Haig & Schnell 2019). It also aims at shedding light on how speakers choose between noun phrases, pronouns and zero arguments, i.e. on which factors influence their referential choice, and it compares these results to other languages.

In the next section, I will give an overview of the theoretical background, i.e. how (radical) pro-drop has been discussed in the literature, how the terms PRO-DROP and RADICAL PRO-DROP came into being and

I will provide definitions of the most important notions. A classification of different types of PRO-DROP will be given. I will also sum up what might influence referential choice according to different studies. In Section 3, I will introduce my research questions and hypotheses, then turn to the data I used and collected, give an overview of Multi-CAST (Haig & Schnell 2019) and therein contained languages, and then explain how I included Mandarin into the corpus and which methodological choices I had to make during that process. I will also explain which quantitative methods I use to analyse the data available to me. The results are then described in Section 4. Namely, I first count the frequency and rate of zero arguments in Mandarin and then compare the results to the other languages. I then use decision trees to predict referential choice in Mandarin according to certain variables and compare these trees to the other languages in the corpus. The results are discussed in Section 5 and put into perspective with respect to current literature. I give a conclusion and examine problems and questions that my study has uncovered and on which further research could focus (Section 6).

2 | Theoretical background

In this section, I will give a short overview of the theoretical background of pro-drop, radical pro-drop and referential choice. I will first sum up the research history on pro-drop, its definition and its different classifications. Since a review of all the literature available on this topic, especially in the domain of Generative Grammar, would go far beyond the scope of this thesis, only the most important observations and claims will be summarised. A special focus will be given to radical pro-drop, which is claimed to be a feature of Mandarin Chinese. Thus I will dive more deeply into radical pro-drop and give a short overview of its definition, of the claims made about radical pro-drop in the literature, and explain how Mandarin fits into the picture.

I will then concentrate on a different but related question, namely referential choice. I will give an overview of its research history and variables claimed to influence it in different languages. In the end, I will turn to Mandarin and give an overview of studies and claims on referential choice and its variables specifically in Mandarin.

2.1 Pro-drop

PRO-DROP refers to the covert realization of a core argument and the distinction between languages that (routinely) allow the omission of ar-

guments, and languages in which the omission of an argument is usually ungrammatical.

For instance, subjects[1] are regularly omitted in Spanish (1), while the same construction would be ungrammatical in English[2] (2) (Roberts & Holmberg 2009: 4, Huang 1984: 532):

(1) Habla español. (Roberts & Holmberg 2009: 4)

(2) * Speaks English.(Roberts & Holmberg 2009: 4)

The first to make this distinction in Generative Grammar was Perlmutter (1971) albeit only for subjects (Roberts & Holmberg 2009: 3f.). He proposed that languages can be classified into 'Type A' and 'Type B' languages (Perlmutter 1971: 115), depending on their permission of zero subjects, and that there was a correlation of this property of a language with other properties, e.g. *that* trace effects and WH-movement, explained below.

The term 'pro-drop' to describe this phenomenon was first coined in Chomsky's (1981) Government and Binding Theory (Ackema et al. 2006: 2, Barbosa 2011b). Subsequently, it became a highly-debated topic in Generative Studies, more precisely within the framework of Principles and Parameters (e.g. Wratil 2011, Sessarego & Gutiérrez-Rexach 2017, Barbosa 2009, Speas 2006, Koeneman 2006, Bennis 2006, Adams 1987, Barbosa 2011a; see Battistella 1985, Huang 1984, Huang 1992 and Liu 2014 for their discussions of Chinese). This framework assumes that there are certain universal principles of Universal Grammar (UG) which

[1]Note that in Spanish, only the subject may be omitted, while the object is obligatorily overt (Huang 1984: 532, Liu 2014: 4).

[2]Note that it is possible or obligatory to omit the subject in certain English constructions (Huang 1984: 532).

Table 1: Rich verbal inflection in Italian versus low verbal inflection in English, adapted from Ackema & Neeleman (2007: 82)

	Italian	*English*
1SG	parl-*o*	speak
2SG	parl-*i*	speak
3SG	parl-*a*	speak-*s*
1PL	parl-*iamo*	speak
2PL	parl-*ate*	speak
3PL	parl-*ano*	speak

represent the innate grammatical knowledge of a child and make it possible for it to acquire language as fast as it does (Wratil 2011: 47). These principles are claimed to vary depending on certain parameters, the pro-drop or null subject parameter being one of these (Chomsky 1981: 231-284, Wratil 2011: 47, Koeneman 2006: 76; see contrary views in Bennis 2006: 101f.).

The stereotypically analysed pro-drop languages Italian and Spanish show rich verbal inflection whereas the stereotypically analysed non-pro-drop languages English and French do not (see Table 1 for a comparison of Italian and English). Thus the (dropped) subject is coreferenced on the verb in Italian and Spanish, but not in English and French. This has led Generative Grammar to claim that rich verbal inflection plays a role in pro-drop (Ackema & Neeleman 2007: 82, Koeneman 2006: 76f., Bennis 2006: 101, Fuß 2011: 53, Huang 1984: 534, Huang 1992: 9, Neeleman & Szendröi 2007: 671, Liu 2014: 3, Travis & Cacoullos 2012: 733). This hypothesis was underlined by diachronic studies, e.g. on the loss of pro-drop in Old French coinciding with the loss of inflectional endings on the verb (Ackema & Neeleman 2007: 82, Wratil 2011: 103ff, see e.g Fuß 2011 for a more critical diachronic study).

However, subsequent research revealed that languages can allow the

dropping of both subject and object even though the verb shows no or little agreement with either (e.g. Ackema & Neeleman 2007 for Early Modern Dutch). The most prominent example of this so-called **radical or discourse pro-drop** (Ackema & Neeleman 2007: 83, Ackema et al. 2006: 5, Barbosa 2011a, Liu 2014) is Mandarin, but it is not the only language displaying this feature (Ackema & Neeleman 2007: 83), which poses a significant problem for the hypothesis that pro-drop and rich agreement systems go hand in hand (Huang 1984: 537, Huang 1992: 9, Neeleman & Szendröi 2007: 672).[3]

Various other characteristics are claimed to be connected to the pro-drop parameter, namely free subject inversion, WH-movement and so-called *that* trace effects (e.g. White 1985, Bennis 2006: 102). Free subject inversion refers to the phenomenon where subjects can occur postverbally as well as preverbally, e.g. in Italian, as in (3) (Roberts & Holmberg 2009: 16).

(3) Hanno telefonato molti studenti. (Roberts & Holmberg 2009: 16)

The *that* trace effect refers to the fact that sentences like (4) are ungrammatical in English, whereas they are not in null subject languages:

(4) * Who did you say that ø wrote this book? (Roberts & Holmberg 2009: 17)

However, the correlations between pro-drop and these two phenomena have been questioned (Bennis 2006: 102). With time, studies on pro-drop showed that there are more nuanced differences between languages regarding pro-drop, e.g. what kinds of arguments can be dropped, the

[3]See also Fuß (2011) for a critical view from a diachronic perspective.

amount of verbal inflection co-referencing the argument, and in which constructions arguments can be dropped. For instance, a distinction must be drawn with regard to which core arguments may be dropped in a language: There are referential core arguments, and non-referential ones.

(5) It rains.

(6) He eats.

In (5), the pronoun is non-referential as it does not refer to any specific entity, whereas the pronoun in (6) refers to a specific human entity, namely the person that is eating. Some languages allow both non-referential and referential arguments to be dropped (e.g. Mandarin), while others only allow non-referential ones to be dropped (e.g. Finnish) (Ackema et al. 2006: 12).

This has led to a distinction between different types of pro-drop languages, which are listed below, sorted according to their scale of freedom in allowing pro-drop[4]:

1. *Non-null subject languages*, e.g. English,
2. *Expletive or semi-null subject languages* (Roberts & Holmberg 2009: 8). This corresponds to the distinction between referential and non-referential arguments made above. In some non-pro-drop languages that generally do not allow omission of subject, it is possible to omit a non-referential expletive subject as in (5).
3. *Partial null subject languages* (Roberts & Holmberg 2009: 10ff., Rosenkvist 2010, Koeneman 2006, Koeneman 2006: 77), namely

[4]This hierarchy is adapted from Roberts & Holmberg (2009: 12).

languages in which pronouns may be omitted, but only under certain conditions, e.g. only the first and second person.

4. *Consistent null subject languages* (Roberts & Holmberg 2009: 6), which are historically the first languages claimed to be pro-drop languages and the languages which have received the most attention. The subject can be omitted in all tenses and persons, and verbal inflection is usually rich. Consistent null subject languages typically also exhibit the above-discussed free subject inversion and *that* trace effects (Roberts & Holmberg 2009: 16, White 1985, Bennis 2006: 102).

5. *Discourse* or *radical pro-drop languages* (Roberts & Holmberg 2009: 8ff, Ackema & Neeleman 2007: 83 Ackema et al. 2006: 5, Barbosa 2011a, Neeleman & Szendröi 2007, Liu 2014) are languages that freely allow pro-drop (namely in all constructions and in all syntactic functions) but do not exhibit rich verbal inflection.

In recent years, there have been new developments corresponding to the availability of large amounts of corpus data, e.g. Multi-CAST (Haig & Schnell 2019) and the advancement of statistical methods in linguistics. This makes it possible to conduct studies on the rate of overt and covert anaphora in different languages (e.g. Bickel 2003, Stoll & Bickel 2009), and even to conduct probabilistic analyses of referential choice in natural language use (e.g. Torres Cacoullos & Travis 2019, Schiborr 2018, Schnell & Barth 2018, Travis & Cacoullos 2012). Bresnan et al. (2005: 2) note that

> [t]heoretical linguists have traditionally relied on linguistic intuitions such as grammaticality judgments for their data.

> But the massive growth of computer-readable texts and recordings, the availability of cheaper, more powerful computers and software, and the development of new probabilistic models for language have now made the spontaneous use of language in natural settings a rich and easily accessible alternative source of data. (Bresnan et al. 2005: 2)

I will now turn to *radical pro-drop* and explain how this notion is connected to Mandarin. I will give an overview of claims about Mandarin in the literature, outline why it is believed to be extraordinary with regard to pro-drop and note where research gaps are to be filled.

2.2 Radical pro-drop

Mandarin has played a prominent role in research on pro-drop languages and is one of the typical examples of so-called *discourse or radical pro-drop* languages (Roberts & Holmberg 2009: 9, Neeleman & Szendröi 2007, Liu 2014). This is due to two claims about the radicality of pro-drop in Mandarin, which will be demonstrated in the next two sections (2.2.1 and 2.2.2).

2.2.1 Free distribution of zero arguments

The distribution of zero arguments in Mandarin is very free even though there is no verbal agreement with any core arguments: it not only includes subjects – which remain the main emphasis of research on pro-drop to date – but also arguments in any other syntactic function, e.g. objects (Battistella 1985: 324, Roberts & Holmberg 2009: 9, Huang 1984: 533, Neeleman & Szendröi 2007: 672, Liu 2014), as illustrated in Examples (7) and (8) below.

(7) *ø kanjian ta le*
ZERO see 3SG ASP
'He saw him.' (Roberts & Holmberg 2009: 9)

(8) *ta kanjian ø le*
3SG see ZERO ASP
'He saw him.' (Roberts & Holmberg 2009: 9)

Interestingly, Huang (1984) notes that in radical pro-drop languages there seem to be other mechanisms and constraints at work to identify reference. This refers to what he calls "subject-object asymmetry", and describes the fact that objects in Mandarin dependent clauses are not bound in reference to the matrix clause, but to the discourse context, while Mandarin subjects and English subjects AND objects are bound to the matrix clause (Huang 1984: 541).

(9) Speaker A:

shei kanjian le Zhangsan?
who see ASP Zhangsan

'Who saw Zhangsan?' (Huang 1984: 539)

(10) Speaker B:

Zhangsan shuo Lisi kanjian le ø
Zhangsan say Lisi see ASP ZERO

'Zhangsan said Lisi saw him.' (Huang 1984: 539)

The claim Huang (1984: 539) makes here is that the reference of the omitted object in (10) is not automatically Zhangsan if the sentence is uttered out of context; rather, a hearer would prefer a reading where it is NOT Zhangsan. Only the discourse context (Example 9) makes a reading possible where the referent of the omitted object is Zhangsan. Comparing this to English, we see that the reference of an object pronoun is bound to the matrix clause:

(11) John said that Bill knew him. (Huang 1984: 538)

In Example (11), the pronoun is interpreted as referring to John if the sentence is uttered without discourse context (Huang 1984: 539).

Note that in my view, discourse plays a role in English as well as in Mandarin, since the pronoun in both languages could have a number of different referents depending on discourse context. In addition, one should be cautious when relying on discourse-free utterances, since these are highly unnatural and thus do not represent actual language use. This is also one of the critical points that Yan Huang (1992) makes against James Huang (1984): "there is, in my opinion, no such thing as a pragmatically neutral linguistic example, since we understand the meaning of a linguistic example only against a set of background assumptions" (Huang 1992: 23).

Yet, he agrees that zero arguments in Mandarin are ultimately "not grammatically but pragmatically determined" (Huang 1992: 27), compared to languages like English, in which zero anaphors are grammatically determined by sentence structure. This raises the question what pragmatic factors influence referential choice, which will be taken up in the next section.

Some researchers have claimed that the pragmatic factors influencing referential choice might differ between pro-drop types. In their view, there is a fundamental difference between pro-drop languages and radical pro-drop languages with regard to referential choice. For instance, because of the poor verbal inflection mechanisms, other mechanisms than co-reference with the verb might determine reference, and other influences might play a role in the choice between noun phrase, pronoun and zero. For instance, Li & Bayley (2018: 137) note that referential choice

in Mandarin may be governed by other factors than in other languages.

2.2.2 Frequency of zero arguments

The comparison of the rate of zero arguments in languages is called referential density by Bickel (2003: 708). Based on renarrations of the pear stories Chafe 1980, Bickel (2003: 708) analysed and compared Belhare, Nepali and Maithili, and found "a statistically significant difference between referential density means in narratives across speakers of different languages" (Bickel 2003: 732).

With regard to Mandarin, there is a persistent claim that zero arguments are very frequent (Li & Thompson 1979, Huang 2000, Yang et al. 2003: 287). For instance, Battistella (1985: 324) claims that they are a "pervasive feature of Chinese", and Bickel (2003: 708) notes that "Chinese discourse [...] is well known for often being very implicit about referents compared to other pro-drop languages". Similarly, Pu (1997: 281) writes that

> [u]nlike English which uses anaphoric pronouns extensively and zero anaphora in syntactically more constrained circumstances, Chinese makes a much lesser use of lexical pronouns in tracking reference and a principal use of zero anaphora in discourse.

This has led some researchers to even claim that languages like Mandarin are fundamentally different from other languages, as pragmatics plays a much larger role in grammar than in other languages. Huang (2000: 261-277) claims that there are "pragmatic languages" and includes Mandarin in this list.

This claim about Mandarin has not been tested quantitatively in a large-scale comparative study, with only a few recent studies that concentrate on probabilistic analyses of referential choice (e.g. Li & Bayley 2018, Pu 1997, Li 2012, Pu 1995) and has mostly been based on intuitive claims on Mandarin made in Generative Grammar.

In the next section, I will give an overview of referential choice, factors claimed to influence it and studies on referential choice in Mandarin.

2.3 Referential choice

As noted in the section above, a related question on radical pro-drop is how referential choice is distributed in discourse, namely when a speaker chooses to use a lexical noun phrase, a pronoun or a zero argument.

There are two different approaches to pro-drop, namely the generative parametric approach discussed above, and the usage-based approach taken by e.g. Bickel (2003), and Schnell & Barth (2018: 58).

The usage-based approach tries to generalise rules about pro-drop based on corpus linguistics (Schnell & Barth 2018: 58), which is the approach I am choosing in this thesis. This approach has the advantage that it analyses actual spoken language within its discourse context, while the generative parametric approach relies on intuition and individual utterances taken out of context. This is often not suitable for explaining a speaker's subconscious probabilistic choice, especially when more than one variable makes an impact and when choices are depending on the discourse context. The study of referential choice is related to the speaker's choice between the three possible forms an argument can take:

> When a speaker is using language, grammar cues him/her to particular choices: which word order to use, where to place

> a discourse marker and which one, etc. Thus grammar is actually a system that guides a speaker's choices. Some choices are relatively rule-based, whereas other choices are rather probabilistic. Discourse-related choices mostly belong to the latter kind: a certain option is not strictly required or strictly ruled out, and more than one option is to a certain extent permissible. (Kibrik 2011: 15)

This is also the case with referential choice; in Mandarin, none of the three forms (NP, pro, zero) would render a sentence ungrammatical; rather, different forms are more or less appropriate depending on context. While understanding the probabilistic rules behind this would be very interesting, they are on the other hand very hard to research, since one cannot rely on introspection to determine these rules. Thus the aim of this thesis is to profit from recent developments in corpus linguistics and statistical advances in linguistics and use probabilistic methods to predict which form a speaker is most likely to use.

This is connected to claims that referential choice is influenced by different factors in different languages, possibly corresponding to the above already-mentioned different classifications of pro-drop. For instance, Ackema et al. (2006: 16) write that "[...] it turns out that the syntax of subjects in pro-drop languages deviates from that of subjects in non-pro-drop languages in a number of respects." Specifically, it has been claimed that Mandarin, as a radical pro-drop language, might act differently to other languages regarding referential choice. Some researchers have suggested that this difference is due to the different verbal marking in these languages: Ackema & Neeleman (2007) claim that since Early Modern Dutch has little verb agreement that could grammatically refer to the dropped argument, pragmatic conditions play a larger role in de-

termining the referential choice of either pronoun or zero in Early Modern Dutch. The dropped argument must then be more salient in discourse than in Italian-style pro-drop languages, the analysis of which Ackema & Neeleman (2007) base on Accessibility Theory (see Section 2.3.1.5).

Ackema et al. (2006: 15) claim that in Mandarin, a dropped argument is the topic of the discourse. They believe that a difference must be drawn between languages with rich verbal inflection that allow pro-drop when the argument remains identifiable through coreference on the verb ('pro-drop'), and languages that only allow pro-drop when the argument is topic and can be identifiable through the discourse context ('topic-drop') (Ackema et al. 2006: 15).

Many studies, however, point in the direction that all languages have the same constraints, e.g. Pu (1995) for English and Mandarin, or Torres Cacoullos & Travis (2019) for English and Spanish:

> Rates of use are not a reliable comparison measure. Despite the conspicuous rarity of unexpressed subjects in English compared with Spanish, there is structured variability within this non-null subject language, which, contrary to cherished belief, displays striking parallels with variation patterns in the null-subject language. (Torres Cacoullos & Travis 2019: 682)

In the remainder of this section, I will give an overview of factors claimed to have an impact on referential choice in the literature. This will then be the basis for the variables I will look at in my analysis of Mandarin.

2.3.1 Factors influencing referential choice

There are several variables claimed to influence referential choice. Some of these are dependent on the discourse context, e.g. antecedent distance, while others are independent of the discourse context, as these are inherent properties of the referent, e.g. animacy.

Since too many factors have been proposed in the literature to include in this description, I limit myself to the most important ones. Factors that have been mentioned in the literature but will not be discussed here, include discourse segmentation (Giora & Lee 1996: 114), definiteness (Ariel 1996: 22), constructions with specific semantic verb classes (Travis & Cacoullos 2012: 725), and backgrounded tense-aspect-mood in Spanish (Travis & Cacoullos 2012: 725).

2.3.1.1 Syntactic Function

According to Du Bois (1987) and Du Bois (2003), information is distributed in discourse in an ergative pattern, and this distribution is what gives rise to ergative languages. Du Bois (2003: 34) posits four constraints on Preferred Argument Structure:

1. Avoid more than one lexical core argument.
2. Avoid lexical A. (see also Du Bois 1987: 823)
3. Avoid more than one new core argument. (see also Du Bois 1987: 826)
4. Avoid new A. (see also Du Bois 1987: 827)

This would imply that the syntactic function of the referent plays a role in referential choice, with transitive subjects rather being pronominal or zero than lexical full noun phrases.

Du Bois (1987: 829) believes that these effects are effects of topic continuity, since the topic is prototypically human, in the A role, and given. This claim thus correlates with the factors of animacy and topicality, discussed below.

Recently, Haig & Schnell (2016) showed that while the A role does have a low rate of full noun phrases, this can be better explained with the variable of animacy, thereby questioning the claims of Du Bois (1987) and Du Bois (2003). This variable will be discussed in the next section.

2.3.1.2 Animacy

Roughly speaking, animacy denotes the distinction between human, animate and inanimate entities (Dahl & Fraurud 1996). In this thesis, animacy as a variable is only concerned with the binary distinction between human and non-human referents. It is an inherent property of the referent and independent of the discourse context, and has been claimed to play a role in referential choice, e.g. by Fraurud (1996), Ariel (1996: 22) and Hsiao et al. (2014).

Hsiao et al. (2014) showed that subject omissions in Mandarin are higher when both the subject and the object are animate, while they are lower when the object is inanimate. Out of the 3810 transitive clauses analysed in their study, 2445 had an overt subject, and 1365 contained a null subject (Hsiao et al. 2014: 4).

In contrast, Schnell & Barth (2018) found that animacy could not correctly predict the choice between pronominal and zero objects. The authors used texts from different registers, which enabled them to compare animate and inanimate discourse topics, and found that discourse topicality, rather than animacy, plays a role in referential choice. The connection between animacy and topicality was also drawn by Pu (1997):

> Among various semantic, pragmatic and discourse factors, animacy (+/- HUM) seems to strongly affect the syntactic coding of a referent because in narrative discourse, a referent that is human is more often topical, agentive, given and definite than a non-human referent, and is more likely to be coded by grammatical subject and hence zero anaphora. (Pu 1997: 290)

In line with the study of Schnell & Barth (2018), Pu (1997: 290) found that animacy (+/-HUM) increases the likelihood of pronouns in contrast to zeros in both English and Mandarin, but noted that this effect was especially high when the referent was topical , while it was low when the referent was not topical.

2.3.1.3 Topicality

As was already shown in the section above, topicality is a widespread notion in information structure, and has been claimed to play an important role in referential choice (Ariel 1996: 22, Schnell & Barth 2018: 73, Huang 1984: 541). However, its definitions vary strongly.

Most importantly, topicality of a referent can either refer to the discourse topic, namely the topic of a certain narrative and "that discourse entity that an entire text is about and that makes the text interesting" (Schnell & Barth 2018: 59), or to the sentence topic, which stays within the boundaries of a sentence, and is often discussed in connection with so-called *topic-prominent* languages, in which topics are claimed to be an important feature of the grammar and which might even have topics (and comments) rather than subjects (and predicates) as their most basic clause structure (e.g. Li & Thompson 1981: 15f., 85ff.).

These two different kinds of topicality need to be kept distinct when discussing its influence on referential choice:

> This suggests that discourse topicality and sentence topicality do not converge when it comes to the use of pronouns for objects in Vera'a, and discourse topicality is a factor in its own right, distinct from the pragmatic relation of topic within a sentence. (Schnell & Barth 2018: 73)

Schnell & Barth (2018) found that discourse topicality was the best predictor for the choice between zero and pronoun in Vera'a. Discourse topics in Vera'a are more likely to be realised pronominally (Schnell & Barth 2018: 59). However, since discourse topicality is not explicitly annotated in Graid (Haig & Schnell 2014) and RefIND (Schiborr et al. 2018), it cannot be analysed in this thesis directly.

An important question thus is how to analyse and define topicality in the corpus. One possibility would be to assume that all human referents are topics, as was done in Schnell & Barth (2018: 59): "for narratives, we assume that all human or human-like referents [...] are the discourse topics, [....]. And for the two types of descriptive texts, it is the fish or plant species, respectively." Accordingly, *animacy* could be analysed as being an indirect indicator of topicality (1). A second analysis of topicality could propose that referents are more topical when they have been mentioned more recently. This would correspond to *antecedent distance* (2). Finally, one could count the most frequent referents in each text and assign them topic status, which corresponds to *overall frequency of referents* (3). These three variables are assumed to indirectly correlate with the topicality of referents in this study.

2.3.1.4 Person

Another factor claimed to influence referential choice is person (Wratil 2011: 119). This is an inherent property of the referent and independent of the discourse context.

In some so-called partial pro-drop languages (see Section 3), e.g. Finnish and Hebrew, only the first and second person pronouns can be omitted (Koeneman 2006: 100). In Vera'a, Schnell & Barth (2018: 74) found that while there was variation between pronominal and zero forms in the third person, all first and second person mentions were pronominal.

Wratil (2011: 119) believes that first person referents are more "topic-worthy" than second person referents, which are in turn more "topic-worthy" than third person referents. The more "topic-worthy" a referent is, the more likely it is to be realised as an unmarked argument, in the syntactic function of a subject, and to have the semantic role of an agent (Wratil 2011: 119). This is because first and second person referents refer to the actual speech act participants (Wratil 2011: 119). Similarly, Ariel (1996: 22) believes that speaker and addressee are inherently more accessible.

Null and overt pronouns in Mandarin carry different information, since the pronouns in Mandarin carry information on person and gender (see also Gelormini-Lezama 2018: 387). It would thus be possible that person affects referential choice between pronouns and zero. Note that full lexical noun phrases are always third person, and should thus be excluded from any analysis of person as an influencing factor.

Li & Bayley (2018: 149) found that person and number played a role in their study of subject omissions in Mandarin. Also, Li (2012: 107) found that singular subjects are more likely to be pronominal, while plural subjects in Mandarin tend to be covert.

2.3.1.5 Antecedent-related factors

Antecedent-related factors are all dependent on the discourse context because they do not concentrate on the anaphoric form itself but rather on its antecedent, namely the last mention of the referent in discourse.

There are several variables that could be tested in this area, but in order to adhere to the limited scope of my thesis, I will explore the in my view most important one, namely antecedent distance, which is also connected to topicality as discussed above.

It might prove fruitful to test other variables in the future, i.e. the syntactic function of the antecedent (Gelormini-Lezama 2018: 387, Schnell & Barth 2018, Travis & Cacoullos 2012), the referential form of the antecedent (Schnell & Barth 2018, Travis & Cacoullos 2012, Torres Cacoullos & Travis 2019), and competition between potential referents (Ariel 1988: 65, Travis & Cacoullos 2012, Li 2012: 107).

The distance to the last mention of a referent is claimed to have an impact on referential choice (e.g. Ariel 1996: 22, Ariel 1988: 65). This factor is closely connected to Accessibility Theory.

Accessibility Theory assumes that referential choice is determined by what the speaker assumes to be the "degree of accessibility of the mental entity for the addressee" (Ariel 1996: 20). The accessibility of a referent positively correlates with the degree of markedness.

Accessibility Theory assumes that speakers and hearers carry mental representations of referents, which can be more or less accessible depending on certain factors, and are marked differently by the speaker depending on their accessibility status (Ariel 1988: 80).

The way mental representations are marked is defined in the *Accessibility Marking Scale*:

(12) The Accessibility Marking Scale, taken from Ariel (1996: 30):
zero < reflexives < agreement markers < cliticized pronouns < unstressed pronouns < stressed pronouns < stressed pronouns + gesture < proximal demonstrative (+NP) < distal demonstrative (+NP) < proximal demonstrative (+NP) + modifier < distal demonstrative (+NP) + modifier < first name < last name < short definite description < long definite description < full name < full name + modifier

With regard to Mandarin, then, the zero argument would be the highest accessible one (Giora & Lee 1996: 113). Ariel (1996: 22) believes that the more recently a referent has been mentioned in discourse, the more accessible it is. This is one of the most central claims of Accessibility Theory, and has been supported by Travis & Cacoullos (2012: 729), who believe that their finding of the relevance of switch reference in referential choice proves the usefulness of Accessibility Theory.

Contrary to these claims, Schnell & Barth (2018) found in their study on referential choice in Vera'a that antecedent distance did not play any role in the choice between pronoun and zero (but Accessibility Theory might still be useful for lexical noun phrases).

Schnell & Barth's (2018) quantitative probabilistic study on referential choice only between pronoun and zero in Vera'a thus does not lend any support to Accessibility Theory; in fact, all variables that Accessibility Theory would predict to have an effect (antecedent distance, discourse interruptions) did not prove to be relevant (Schnell & Barth 2018: 69).

> Finally, our findings provide ample counterevidence to the universal relevance of accessibility and activation, suggesting

> that at least the choice between pronoun and zero for objects is not accountable for in terms of AT and similar frameworks concerned with discourse structure. From an activation point of view, these two forms of reference appear to be too similar to mark significant differences. (Schnell & Barth 2018: 76)

Consequently, accessibility theory and antecedent distance might play an important role in the distinction between lexical noun phrases and pronominal / zero arguments, while they seem to be irrelevant for the distinction between pronominal arguments and zero arguments.

2.3.2 Referential choice in Mandarin

It has long been claimed, though on a purely intuitive basis, that Mandarin behaves fundamentally differently from other languages, the following quotation being a response to these claims:

> We recognize that speakers' choices between overt and null pronouns are likely to pattern differently in a radical pro-drop language like Chinese, which lacks verbal inflections to indicate person and number, than in an inflected language like Spanish. (Li & Bayley 2018: 137)

For a long time, these claims had not been tested, but in recent years, there have been more and more studies on referential choice and probabilistic analyses outside the realm of GG, focusing on corpus studies. Some of these have circled around Mandarin as a radical pro-drop language, e.g. Li & Bayley (2018), Pu (1997), Li (2012), Pu (1995). Pu (1995) analyses anaphoric distribution in Mandarin and compares it to English, concluding that they are subject to the same constraints in

anaphoric distribution in narratives (narrative production task, Pu 1995: 280).

Pu (1997: 286) investigates pragmatic, semantic and discourse actors in the distribution of zero anaphora, using the first 25 pages of three contemporary Chinese novels. Li (2012) analysed speech from three different discourse contexts and used logistic regression (Li 2012: 102) to analyse subject pronominal expression. They found that switch reference, person, number, animacy, specificity and sentence type played a role, as well as sociolinguistic factors of the speakers.

2.4 Interim conclusion

In this chapter, I have given an overview of pro-drop, radical pro-drop and referential choice in general, and in Mandarin, specifically.

I have shown that languages can be grouped into different classifications with regard to the grammaticality and ungrammaticality of zero arguments in different constructions and positions. The most important types for the purpose of this thesis are

1. non-pro-drop languages, e.g. English,
2. pro-drop languages, e.g. Spanish, and
3. radical pro-drop languages, e.g. Mandarin.

While zero arguments are mostly ungrammatical in a simple declarative sentence in English, and can only occur in specific constructions that grammatically force the omission of an argument or when the subjects are co-referential, zero arguments are generally omitted and grammatical in Spanish and Mandarin. Yet, in Spanish, these zero arguments are

limited to subjects, while all arguments can freely be omitted in Mandarin. In addition, Spanish exhibits a rich verbal agreement system, while Mandarin does not co-reference its core arguments on the verb.

Mandarin has played an important role in the literature (e.g. Battistella 1985, Huang 1984, Huang 1992, Liu 2014, Roberts & Holmberg 2009, Neeleman & Szendröi 2007) because a) it was claimed that it freely admits the omission of all arguments (e.g. Battistella 1985: 324, Roberts & Holmberg 2009: 9, Huang 1984: 533, Neeleman & Szendröi 2007: 672, Liu 2014), and b) that these zero arguments are very frequent, even compared to other pro-drop languages (i.e. Li & Thompson 1979: 322, Huang 2000: 262, Yang et al. 2003: 287, Battistella 1985: 324, Bickel 2003: 708, Pu 1997: 281).

While these claims persist to this day, there has been little quantitative research to test them and compare Mandarin to other languages.

Referential choice is concerned with the question of how the speaker chooses which form to use in discourse when all three are grammatically possible: the full lexical noun phrase, the pronoun, or a zero argument. There are persistent claims that Mandarin as a radical pro-drop language acts more pragmatically than other languages and that choices in referential choice differ from choices in other languages. These claims have rarely been tested, and more quantitative studies on this are needed (see e.g. Li 2012: 116 and Travis & Cacoullos 2012: 743):

> The study of referent realization can be advanced through the pursuit of accountable quantitative studies, in different language varieties; taking into consideration different genres, persons and syntactic roles; employing replicable operationalizations of notions to be tested; exploring further the workings of accessibility and the strength and interactions of priming

> effects; and identifying fixed constructions which may exhibit distinct behavior. (Travis & Cacoullos 2012: 743)

In order to test these claims in the subsequent chapters, I gave an overview of potential drivers of anaphoric distribution discussed in the literature. Among these are syntactic function, topicality (= animacy, antecedent distance, overall frequency of referents), and person. Some of these factors correlate with each other. For instance, referents are often at the same time subjects, human, topical, and have low antecedent distance. A statistical analysis should thus be suitable for this kind of data and methods that can account for correlating variables should be chosen.

It was also shown that referential choice between noun phrases, pronouns and zero arguments might be due to different factors than the distribution of pronoun and zero. This was shown in Schnell & Barth (2018), where antecedent distance seemed to play a role in the choice between noun phrases and non-lexical arguments, but not in the choice between pronoun and zero; but also when looking at the factor of person, which does not play a role for noun phrases, since they are always in the third person.

Based on the above-discussed research gaps and claims about pro-drop in Mandarin, I will now formulate the research questions and hypotheses, and then describe the corpus I will use to test the hypotheses. As the collection and preparation of the Mandarin sub-corpus has been part of this thesis, I will discuss it in greater detail, i.e. my methods of collecting the data, and any methodological decisions I had to make while annotating the data. I will then clarify the methods I use in the quantitative analysis, how it fits with the correlating variables, and some choices I made regarding the exclusion and inclusion of certain data points.

3 | Methods

In this chapter, I will present the research questions and hypotheses. Then, I will give an overview of the data that I will be using for the quantitative study, namely the Multi-CAST corpus (Haig & Schnell 2019), in which Mandarin has been included as part of this thesis.

After a summary of all the languages available in the corpus at the time of analysis, I will describe the quantitative analysis and explain the methods, software and packages that I use and the variables that I consider for the probabilistic analysis of referential choice.

3.1 Research questions and hypotheses

From the reviewed literature in Chapter 2, two research questions arise that I devote the analysis of this thesis to:

Research Questions

1. Is there a higher rate of zero arguments in Mandarin than in other languages?

2. Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages?

    a) Language

b) Syntactic function (Section 2.3.1.1)

c) Topicality (Section 2.3.1.3)

   i. Animacy (+/-HUM) (Section 2.3.1.2)

   ii. Antecedent distance (Section 2.3.1.5)

   iii. Overall frequency of referents (Section 2.3.1.3)

d) Person (only for the distinction between pronoun and zero argument, Section 2.3.1.4)

The first research question responds to the claim that the rate of zero arguments in Mandarin is higher than in other languages. The second research question aims at investigating if referential choice in Mandarin is determined by other variables than in other languages.

The constraints that I want to test in the second research question are inspired by what has been shown to be relevant in the literature before, and what I have discussed in the previous chapter. The constraint of *language* responds to the question if all languages have the same constraints, or if they differ in their constraints. If Mandarin really is fundamentally different from other languages, *language* will be one of the statistically significant constraints.

Syntactic function responds to the claims made by Du Bois (1987, 2003) but since Haig & Schnell (2016) have shown that this is actually connected to animacy, I expect that it will not turn out statistically significant. Rather, I hypothesise that animacy (+/-HUM) will be the distinguishing factor.

I hypothesise that *topicality* plays a role in influencing referential choice. Next to animacy, antecedent distance and overall frequency of referents are expected to play into referential choice which indirectly shows topicality.

As regards the choice between pronoun and zero, I will also analyse *person*. Since noun phrases are always in the third person, this factor has to be excluded in the analysis of noun phrases.

In conclusion, I formulate the following hypotheses to the two research questions:

1. Speakers of Mandarin do not use a higher rate of zero arguments than speakers of other languages in the corpus.

2. Probabilistic constraints influence referential choice. These constraints are the same in every language in the corpus. Language and syntactic function do not influence referential choice. Topicality (= animacy (+/-HUM), antecedent distance and overall frequency of referents) and person influence referential choice.

In the following section, I will now give an overview of the data I use to test the hypotheses. Afterwards, I will explain the statistical and quantitative methods that I use for the analysis.

3.2 The corpus

The approach I adopted in this thesis is a corpus-based and usage-based one that differs from the approach used in Generative Grammar.

While in earlier linguistics and in Generative Grammar, grammatical rules were often stated in terms of introspection and intuition, I agree with e.g. Chambaz & Desagulier (2016: 3) that intuitions do not always show the full picture, especially with linguistic variation, and that a database of natural language usage can give more information on how speakers actually speak. In linguistics, this kind of database is often a corpus, which is

> a large-scale collection of texts sampled from genuine linguistic productions by native speakers. From a statistical viewpoint, a corpus is a sample drawn from the true, unknown law of a given language. (Chambaz & Desagulier 2016: 1)

The database of natural language usage that I use in the thesis is the *Multilingual Corpus of Annotated Spoken Texts* (Multi-CAST, Haig & Schnell 2019). I will give an overview of this corpus in the next section.

3.2.1 Multi-CAST (Haig & Schnell 2019)

At the time of data analysis, Multi-CAST (Haig & Schnell 2019) consisted of eight sub-corpora, but more languages have been added recently.[5]

The goal of Multi-CAST (Haig & Schnell 2019) was to develop a system of syntactic annotations that is *flexible* enough to be applied to typologically diverse languages, and at the same time still *consistent* enough to enable quantitative cross-linguistic analysis between languages (Haig & Schnell 2018 [2016]: 1).

Zero arguments are usually not added in other corpus annotations, which means that studies on zero arguments have to add these in themselves. This poses a problem for cross-linguistic comparisons, since annotations of zero arguments have to be consistent across languages in order to be comparable. A huge advantage of Multi-CAST (Haig & Schnell 2019) is thus that annotations include zero arguments, with clear and

[5]This might lead to inconsistencies with regard to tier names (e.g. RefLex has been changed to ISNRef), and some languages might have additional information newly available now. The version I chiefly base my analysis on is version 1905 which was the most recent one at the time of analysis; however, I have included preliminary results from later versions (1907 and 1908) where necessary. The corresponding table as well as R-scripts and the Mandarin data can be found in the Appendix for maximum transparency and reproducibility.

strict guidelines on when to add them. It is therefore possible to not only count the rate of zero arguments in the different languages consistently, but also to analyse referential choice without losing one of the most crucial parts of it, namely covert arguments (Haig & Schnell 2018 [2016]: 2).

Multi-CAST (Haig & Schnell 2019) offers several different tiers of analysis in ELAN,[6] namely GRAID, RefIND and RefLEX (as well as standard tiers on morphological glossing, free translation and transcription).

GRAID (Haig & Schnell 2014) annotations provide us with information on the syntactic function, morphological form and animacy features of referential expressions (Haig & Schnell 2014: 2), which is consistent over corpora and languages (Haig & Schnell 2014: 3).

Studies that have used Multi-CAST (Haig & Schnell 2019) in the past include Haig & Schnell (2016), Haig et al. (2017), Schiborr (2018), Kimoto (2018), Schnell & Barth (2018), Schnell et al. (2018), Schnell & Schiborr (2018).

In the following section, I will give an overview of the languages contained in Multi-CAST (Haig & Schnell 2019) that were used in the analysis. An overview of the languages and their geographical distribution can be seen in Figure 1.

[6]ELAN Version 5.2 [Computer software] 2018, April 4, http://tla.mpi.nl/tools/tla-tools/elan/ (Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands), see also Brugman & Russel (2004).

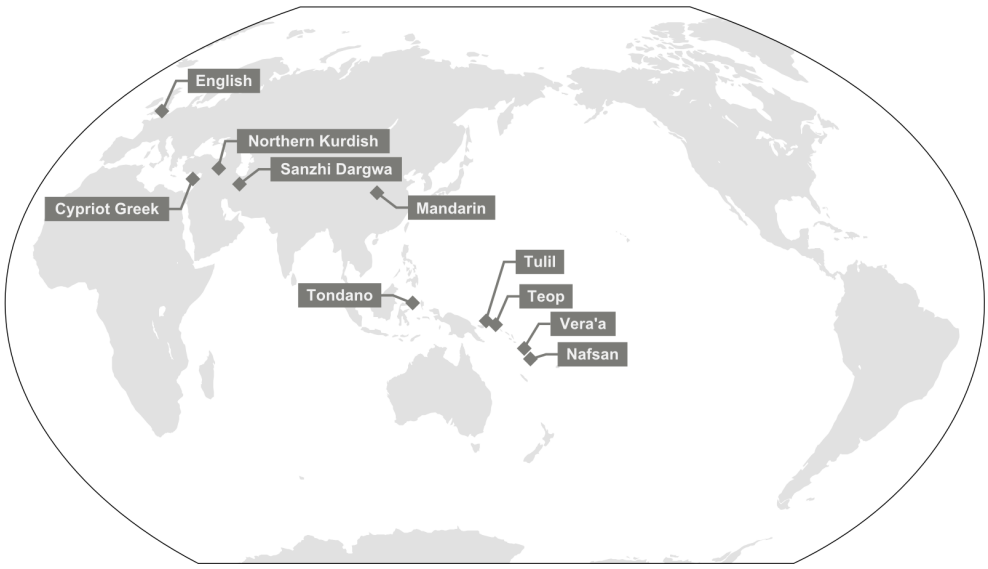

Figure 1: Multi-CAST Languages.[7]

3.2.2 Languages

I have included the following languages in the analysis: Mandarin, Northern Kurdish, Sanzhi, Tondano, English, Cypriot Greek, Vera'a and Teop (see Figure 1). Other languages that may be available in the corpus now could not be used, since they were not available at the time of analysis. I also did not use Persian (Adibifar 2016, 2019), since the corpus consists of renarrations of the *Pear stories* (Chafe 1980, Schiborr 2016), and thus differs a little in this regard from the other languages. With respect to the second research question,[8] only languages that include RefIND could be included in the analysis, namely Mandarin, Cypriot Greek, Sanzhi, Teop and Vera'a. A list of speakers of all languages and their metadata can be found in Schiborr (2016).

[8]Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages?

Northern Kurdish (Haig et al. 2019a) Northern Kurdish belongs to the West Iranian languages (Haig 2018). The Kurmanji corpus consists of traditional narratives (Schiborr 2016: 4).

Sanzhi (Forker & Schiborr 2019) Sanzhi is a Nakh-Daghestanian language spoken in central Daghestan, Russia (Schiborr 2019a: 12). The corpus consists of traditional and autobiographical narratives (Schiborr 2019a: 12).

Tondano (Brickell 2016) Tondano is an Austronesian language spoken in Indonesia (Schiborr 2016: 5). The corpus is made up of autobiographical and stimulus-based narratives, differing in this regard from most other languages in the corpus (Schiborr 2016: 5).

English (Schiborr 2015) English belongs to the Indo-European family (Schiborr 2016: 4). The corpus consists of autobiographical narratives (Schiborr 2016: 4).

Cypriot Greek (Hadjidas & Vollmer 2015) Cypriot Greek belongs to the Indo-European language family (Schiborr 2016: 4). The corpus contains traditional narratives (Schiborr 2016: 4).

Vera'a (Schnell 2015) Vera'a is an Austronesian language spoken on Vanuatu (Schiborr 2016: 5). The corpus consists of traditional narratives (Schiborr 2016: 5).

Teop (Mosel & Schnell 2015) Teop is an Austronesian language spoken in Papua New Guinea (Schiborr 2016: 5). The corpus consists of traditional narratives (Schiborr 2016: 5).

[8]Note that Tulil and Nafsan are not included in the analysis, since they were not available. I am indebted to Nils Schiborr who made this map available to me.

3.2.3 Mandarin

Mandarin belongs to the Sino-Tibetan language family (Li & Thompson 1981: 2). It is isolating (Li & Thompson 1981: 10) and is often described as a topic-prominent language (Li & Thompson 1981: 15, Li & Thompson 1976). The Mandarin data set consists of three monologic, natural narratives from three different native speakers of Mandarin. The texts were translated, transcribed and then annotated with a morphemic gloss, GRAID (Haig & Schnell 2014), RefIND (Schiborr et al. 2018) and RefLEX, respectively. The Mandarin sub-corpus will be published in Multi-CAST after the completion of this thesis.

All stories were told in *Putonghua* (Mandarin), the official national language of China (Li & Thompson 1981: 1). Note that Mandarin in itself is in many ways an artificial construct that is taught to children at school and may still be highly influenced by regional differences between speakers (Li & Thompson 1981: 1). The traditional oral narratives were recorded in Xi'an, China, by the author during an exchange semester in 2015 and 2016. Two of the speakers were originally from northeastern China, and one speaker was from Xi'an. The speakers were my schoolmates or friends. They were asked to tell a story of their choice and were recorded while telling it. Speakers were informed that their stories would be used for research and published online. They were not informed of the specific research questions of this study. It was agreed that they could stay anonymous if they wanted. They were not paid for the recordings, but mostly invited to eat together and spend time with me beforehand. In addition, small presents from Germany were given to them where appropriate. The three stories contain 1175 clause units altogether. More stories have been recorded and transcribed and will hopefully be added to the corpus in the future.

Figure 2: Home of one of the speakers.[9]

3.2.3.1 Jigongzhuan (JGZ)

This speaker is from the northeast of China, he is a university student and was 23 years old at the time of recording. The narrative tells anecdotes of the life of an eccentric Buddhist monk. This story is quite famous in China. Since there was unfortunately no better place, the story was told inside a university building, which means that there is some background noise from time to time that does not disturb the story, however.

The narrative was told in a group of friends and schoolmates, who were all native speakers of Chinese, and every one of them told a story. Since it would have been impolite and distancing to leave the room instead of listening to the story, I was present as well.[10] I still stayed in

[9]Photo: Maria Vollmer, with the permission of the owner of the house.

[10]As a non-native speaker of Chinese, I tried to leave the room whenever possible, in order to avoid possible influences on the way the story would be told when a listener

the background and encouraged the speaker of the story to tell it to his friends rather than to me. The story is 21 minutes and 15 seconds long. It consists of 720 clause units, eight of which are unclassifiable and were thus excluded from analysis.

3.2.3.2 Liangzhu (LZ)

The speaker of this story comes from Shaanxi Province, is a university student and was 22 years old at the time of recording. He tells the romantic love story of a couple that cannot be together because of societal pressure and expectations. The story was recorded in my apartment, since it was comparatively quiet and to make sure we were not disturbed during the recording. The speaker told the story to his friend. I left the room during the recording, thus only native speakers of Chinese were present. The recording is eight minutes and 13 seconds long. The story contains 189 clause units, seven of which were unclassifiable and will not be included in the analysis.

3.2.3.3 Mulan (ML)

The speaker of this story is from the northeast of China, is a university student and was 23 years old at the time of recording. He tells the story of Mulan, a woman that, dressed as a man, secretly went to war in place of her father, and became a war hero. The story was recorded in the same setting as *Liangzhu*, namely in my apartment. I left the room during the recording so that only native speakers were present. The story is ten minutes and 25 seconds long. It consists of 306 clause units, five of which are unclassifiable and have thus been excluded from analysis.

is not a native speaker of Chinese.

3.2.3.4 Corpus annotation

The data were analysed using the software ELAN,[11] and annotated according to Multi-CAST standards, namely the GRAID annotation guidelines (Haig & Schnell 2014) and RefIND annotation guidelines (Schiborr et al. 2018).

The stories were transcribed by Liu Ruoyu as part of her work at the Department of General Linguistics (University of Bamberg), to whom I owe many thanks. She also helped me with questions on the translation or linguistic structure of sentences.[12]

The three stories were annotated using all layers or tiers of Multi-CAST annotations, namely a transcription, a free translation, a morphologic gloss, GRAID, RefIND and RefLex. There were some language-specific choices to be made in the annotations by the author. The most important ones were about serial verb constructions, topic constructions, flexible word classes and the so-called differential object marking, discussed in the paragraphs below in detail. All of these problems were discussed with the Multi-CAST Team (Geoffrey Haig, Stefan Schnell and Nils Schiborr) before being implemented.

Differential Object Marking In Mandarin, the canonical word order SVO (Iemmolo & Arcodia 2014: 316) can be changed when the object is moved in front of the predicate and marked with a preposition, e.g. BA and GEI (Li & Thompson 1981: 463, Liu 2007.

[11] ELAN Version 5.2 [Computer software] 2018, April 4, http://tla.mpi.nl/tools/tla-tools/elan/ (Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands), see also Brugman & Russel (2004).

[12] Of course, many other native speakers helped me whenever I had questions. Unfortunately, it is impossible to mention them all here, but the most important ones are Wu Shuang, Song Jian, Wang Lei and Zhang Jujia.

(13) *ZERO* *jiu* ***gei*** *na* *ge* ***chanpoer*** *xia* *le* *yi*
0_he MP **give** DEM CL **midwife** scare ASP one
0.h:s other **adp** ln_dem ln_cl **np.h:obl** v:pred rv_asp rv
tiao
jump
rv
"He already scared the midwife." (mandarin_jgc_105)

In GRAID, we agreed to gloss preverbal 'objects' that are marked with an adposition as 'OBL' instead of 'P' since, from a strictly formal perspective, these are marked with an adposition and thus not canonically-marked objects. We do not think that this is differential object marking in a narrow sense, but since it is mostly called DOM in the literature (e.g. Iemmolo & Arcodia 2014), this is what we call it here for pragmatic reasons.

Serial verb constructions This construction is problematic in a number of other languages in the corpus as well, e.g. in Northern Kurdish (Haig et al. 2019b). In Mandarin, serial verb constructions are formally very similar to (and often indistinguishable from) topic chains in which multiple predicates occur as a string of verbs and their co-referential argument(s) are covert. While there are various language-specific means of differentiating serial verb constructions from multiple predicates, often involving the scope of negation or TAM markers over the whole predicate instead of one single verb, in practice, most occurrences of serial verb constructions in the corpus are formally ambiguous and thus indistinguishable from topic chains with zero arguments:

(14) *## zhe ge fufu lia ren jiu qu guoqing*
DEM CL couple two person MP go Guoqing
ln ln np.h:a ln np.h:appos other v:pred ln
si ## ZERO bai fo
temple ## 0_they pray Buddha
np:p ## svc_0.h:a v:pred np:p

"The couple went to Guo Qing temple to pray to Buddha." (mandarin_jgz_0065)

(15)

## *ZERO* *zou*	## *ZERO*	*jin*	*le*	*zhe*	*ge*		
## 0_she walk	## 0_she	go.in	ASP	this	CL		
## 0.h:s v:pred	## svc_0.h:a	v:pred	rv	ln	ln		
yingqin	=*de*	*huajiao*					
procession.to.get.bride	=MOD	marriage.sedan					
ln	=ln	np:p					

"to walk in the marriage sedan for the procession [to escort the bride to the bridegroom's home for the wedding]." (mandarin _lz_0100)

In (15), *zou* and *jin* could be interpreted as being one single predicate denoting 'to walk in', but could also be interpreted as two predicates in a topic chain denoting the process of 'to walk' first, and then 'to go in' somewhere. Simply on formal grounds, the second interpretation would be more correct, since there is no formal marking that tells us that the two verbs should be analysed as serial verbs.

In these cases, the constructions are thus glossed as multiple predicates with covert arguments.'svc_' is added to the ZERO gloss. This enables GRAID to capture as much information as possible, while still giving researchers the possibility to exclude these zeros and thereby analyse these constructions as serial verb constructions.

In cases in which

1. the string of verbs clearly denotes a single event or action, or
2. analysing the verbs as multiple predicates in a topic chain would change their meaning in a way that would be contextually incorrect,

the construction is analysed as a serial verb construction. In these cases, the main verb is glossed 'v:pred' and the other verb is glossed 'svc_lv' or 'svc_rv'.

(16) *## ZERO jiù ba ta dài guòlái le*
ZERO ADV ADP 3SG bring come.over ASP
0.h:s other adp pro.h:obl v:pred svc_rv other
"He/they brought him over." (mandarin_jgz_224)

In example (16), we know from context that neither of the participants comes over, as the subject of the clause is already in the right place, and the object of the clause is a new-born baby that could thus not be the subject of *guolai*. Thus this construction is a serial verb construction and *guolai* changes its semantics to a simple directional meaning.

In all the three stories, there are 71 instances of svc_lv/rv and 60 instances of svc_0. For comparison, there are overall 589 zero arguments in the Mandarin sub-corpus. These cases thus make up 22%.

Flexible word classes Mandarin has relatively superfluous word classes. For instance, Sun (2006: 206) notes that "[n]early all Chinese prepositions can be used as full-fledged verbs." With regard to the corpus, this poses a problem for prepositions that also act as verbs and are often used in serial verb constructions. In these cases, the question is if they are to be annotated as verbs or as prepositions; and, if they are analysed as verbs, if they are serial verb constructions or two clauses. This also affects the annotation of the argument after the verb, since it would be object ('p') if analysed as a preposition, but oblique ('obl') if analysed as a verb. An example for this can be seen in (17):

(17) *## ZERO yunyou dao zanmen zheer*
0_he travel reach 1PL.INCL here
##ds 0.h:s v:pred adp ln_pro.1:poss pro:g
"He has traveled to us." (mandarin_jgz_226)

Here, *dao* could also be analysed as a verb, and the clause could then be analysed as two clauses which would also increase the rates of zero arguments. This also means that *zheer* is annotated as being a goal,

while it would be analysed as an object if *dao* were a full verb. However, I have chosen to analyse these instances as prepositions, since this is the primary use of the word, and since this is the analysis in which I presuppose the least and am closest to the actual formal representation.

For comparison, in other cases, *dao* is used alone and as a full verb, as in (18):

(18) *## ZERO dao le dangpu*
0_they reach ASP pawn.shop
0.h:s v:pred rv np:p
"(They) came to the pawnshop." (mandarin_jgz_0427)

Topic constructions Mandarin is claimed to be a topic-prominent language, in which topics are claimed to be an important feature of the grammar and which might even have topics (and comments) rather than subjects (and predicates) as their most basic clause structure (e.g. Li & Thompson 1981: 15f., 85ff.). Topics may be separated from the rest of the clause by pause particles (Li & Thompson 1981: 86), like *ne* in Example (19). Note that here the topic is repeated as a subject in pronominal form:

(19) *er liangshanbo* ***ne*** *ta ziji ye juede*
but Liangshanbo **MP** 3SG REFL also think
other pn_np.h:dt_s_ds **other** pro.h:s_ds other other v:pred
'And Lianshanbo, he himself thought' (mandarin_lz_040)

When subjects are separated from the rest of the clause with a pause particle and the subject is not repeated, as it was the case in Example (19), ZERO is added in the gloss:

(20) *zhe* *ge* *daoji* ***ba*** ***ZERO*** *pingshi* *zai*
DEM CL Daoji **MP** **0_he** usually in
ln ln pn_np.h:dt_a **other** **dt_0.h:a** other adp
siyuan *li* *nian* *nian* *jing*
temple.yard in read read scriptures
np:l adp v:pred rv np:p
'This Daoji, (he) usually read the scriptures in the temple yard.'
(mandarin_jgz_197)

Here, *daoji* is separated from the rest of the clause with the pause particle *ba*[13], and is thus analysed as topic. Since the referent is not repeated overtly as a subject, as in Example (19), ZERO is added in the gloss.

ZERO is not added when the subject is a lexical noun phrase without pause marker (Example 22) even though the subject may still be repeated in the pronominal form (Example 21). The reason for this is that there is no formal marking on the subject which lets us know that it is the topic, except for its leftmost position in the clause.

(21) *danshi* ***zhu*** ***yuanwai*** ***ta*** *you* *yi* *ge*
but **Zhu** **landlord** **3SG** have one CL
other **pn_np.h:dt_a** **rn_np** **pro.h:a** v:pred ln ln
nüer
daughter
np.h:p
'But Zhu landlord, he had a daughter.' (mandarin_lz_010)

(22) *ranhou* *liangshanbo* *jiu* *jueding* *le*
then Liangshanbo ADV decide ASP
other pn_np.h:s other v:pred rv
'Then Liangshanbo decided' (mandarin_lz_072)

It is important to remember that these examples are not as exotic as one might believe; examples like these are abundant in spoken English

[13]This is a different *ba* than the preposition used in DOM.

and any other spoken language as well, and no ZERO would be glossed in these languages either.

When the object is preverbal and in the leftmost position of the clause, it is analysed as the topic of the clause. In this case, a ZERO is added in the gloss (see Example 24), as the object may be repeated in its usual position when it is the topic of the clause (see Example 23), and it would come after the predicate according to canonical word order.

(23) *zhe* *ge* *jidian* *ZERO* *jiu* *xian* *bu*
DEM Jidian CL 0_you MP first NEG
ln ln np.h:dt_p 0.2:a other other other_neg
guan ***ta*** *le*
care.about **3SG** ASP
v:pred **pro.h:p** other_asp
'This Jidian, do not care about him for now.' (jgc_398)

Here, *Jidian* is in the leftmost position of the clause and then realised again after the predicate as a pronoun. It is thus the topic of the clause, while the pronoun is the object. In the next example, the topic is not repeated as the object and ZERO is added in the gloss:

(24) *ni* *de* *laili* *wo* *zhidao* ***ZERO***
2SG MOD origin 1SG know **0_it**
ln_pro.2:poss ln np:dt_p pro.1:a v:pred **dt_0:p**
'Your origin, I know (it).' (jgc_137)

Because these distinctions are somewhat arbitrary and it might be argued that the argument of the verb is in fact overt in the clause, the gloss is extended to 'dt_0', thus making it possible to exclude those ZEROs from analysis or change them in later versions.

In the Mandarin sub-corpus, there are 31 instances of dt_0, which represent only 5% of all zero arguments.

3.3 Quantitative analysis[14]

As discussed above, all arguments in Mandarin can take one of three forms: a full noun phrase, a pronoun and zero.

It has often been assumed that referential choice is determined by the same factors, irrespective of whether it is a noun phrase, pronoun or zero. This can be seen in Ariel's Accessibility Theory (1988, 1996), where all referential forms are ordered along the same continuum, being subject to the same constraints.

However, recent studies have questioned this connection and suggested that the factors at play are different between lexical and non-lexical choice (= noun phrase vs. pronoun / zero) and pronominal or zero argument (e.g. Schnell & Barth 2018 and Schiborr 2018).

Since the answer to these questions has not been answered conclusively, I will concentrate on two perspectives:

1. First, what is the rate and distribution of noun phrases, pronouns and zero arguments?

2. Secondly, what is the frequency and distribution of pronouns and zero arguments, including first and second person?

With regard to the first research question[15], the goal is to simply count the rate of noun phrases, pronouns and zero arguments in the different languages and syntactic functions to see if there is a difference

[14]Quantitative analysis would not have been possible without Jan H. Boockmann (University of Bamberg) who graciously agreed to explain prediction models and approaches in statistics to me, looked at my data and improved my scripts where necessary. Nils Schiborr (University of Bamberg) looked over my script and quantitative analysis many times and I partly used his script for the analysis. I also owe thanks to Stefan Schnell and my supervisor Geoffrey Haig for their feedback on the annotation of the Mandarin sub-corpus as well as quantitative analysis.

[15]Is there a higher rate of zero arguments in Mandarin than in other languages?

in the numbers between Mandarin and the other languages. The analysis is conducted using the statistics software R and RStudio (R Core Team 2019, RStudio Team 2018). The following packages are used for data analysis and data visualisation: Ooms (2019), Schiborr (2019b), Dowle & Srinivasan (2019), Wickham (2016), Neuwirth (2014), Harrell Jr (2019).

With regard to the second research question,[16] I have decided against using multilinear regression, since it does not fit with the nature of my data as it cannot account for correlations between independent variables. As shown above, a number of variables correlate with each other, e.g. animacy, topicality and syntactic function. Since these variables are expected to correlate with each other, a multilinear regression model is thus not the perfect fit for the data (see also Schnell & Barth 2018: 63, Chambaz & Desagulier 2016: 9).

For this reason, I will apply a predicting model instead. In the case of predicting models, there are two different kinds of learning approaches for predicting outcomes. *Black-box learning* (e.g. deep learning) is advantageous as the predicted outcome can be very precise, but a disadvantage is that it does not give any information on how the prediction was made ("*predicting* is not *explaining*" (Chambaz & Desagulier 2016: 9)). *White-box learning* on the other hand has the advantage of explaining which variable feeds into the prediction at what point, but a disadvantage is that variables need to be defined by me and are not found automatically by the algorithm. Since I want to use the prediction approach in order to see which variables play a role and how important they are, a white box learning model was preferred.

For this reason, a decision tree was used (similar to Schnell & Barth 2018), since it makes it possible to simultaneously test variables that

[16] Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages?

correlate with each other, and to see which one of these variables is more important (see also Schnell & Barth 2018: 63). In a decision tree, the algorithm basically iterates over the data to find the variable which has the most impact and is best in explaining the majority of the data, then it divides the data into two subsets based on this variable and continues to recursively scan for the most important variable in both subsets, respectively. This process continues until the algorithm cannot find any impacting variables anymore, or all the data points have been explained. Optionally, the size of the decision tree can be limited to achieve a model that performs slightly worse compared to a decision tree of unlimited size, but is simpler to understand. Since no decision trees in my study are too complex, I did not limit the depth of the trees. This means that when correlating variables are used, the algorithm finds the variable that explains the data best, and the other correlating variables either do not have an impact in the resultant subsets, or still have an impact irrespective of their correlation with the first variable. In either case, the correlations between variables are accounted for and do not distort the results.

The decision trees are completed and visualised using the *rpart* (Therneau & Atkinson 2019) and *rpart.plot* package (Milborrow 2019).

For reasons of time, space and low quantities of data, I do not use two methods that would further ensure the correctness of my analysis, namely a) a *conditional random forest analysis* and b) a division in *training and test subsets*. Even though this study is an important first step, further studies in this area should use these methods to confirm that the analysis in this thesis is correct and test whether results are robust when other methods are applied.

The *conditional random forest analysis* (see also Schnell & Barth

2018: 64) produces several decision trees using subsets of the data, and then completes a decision tree based on what the majority of the trees in the forest has produced. While random forest analysis is a much stronger and more robust method in predicting outcomes, it is harder to control and interpret. Thus, for an explanatory analysis like mine, a decision tree can be advantageous, as long as it is seen as preliminary and its results are compared with other methods (i.e. the barplots and frequencies of referential forms).

I have also not divided the data into a *training set and an unseen test set* like Bresnan et al. (2005: 14), which should be done in future studies, especially when additional data will have been annotated and made available. The division in training and test subsets is similar to the division into subsets in the random forest analysis. It tells us if the tree still works with data that it has not been trained with, thus making sure that the trees function for natural speech in general, and not only for the small sample I use.

Since the amount of data available now is too small to do both the random forest analysis and the division into training and testing subset, the application of both methods will have to wait until more data is published. The results with regard to the decision trees should thus only be seen as preliminary that will have to be confirmed with the help of more-suitable statistical methods, which, however, require a larger dataset.

Nevertheless, they present an important first step towards understanding referential choice and pro-drop in Mandarin based on a sample of natural language use, on which future studies can build.

Language and speaker are directly encoded in Multi-CAST (Haig & Schnell 2019) and can thus be used directly after having been imported

in R. Syntactic function and animacy (+/-HUM) are encoded directly as well. However, there are certain modifications to be made, i.e. with regard to the decision if subjects of a verb of direct speech should be included in the analysis (see Haig & Schnell 2014: 48).

With regard to animacy (+/-HUM), Multi-CAST encodes human participants, anthropomorphised participants and all other participants (Haig & Schnell 2014: 12). The author decided to include anthropomorphised referents into the analysis of human referents. Note that a preliminary analysis showed that human and anthropomorphised referents tend to cluster together, which supports this decision.

Person is annotated directly in the data as well, but can only be included in the analysis when noun phrases are excluded, since these are always in the third person singular.

However, all antecedent-related variables are only available indirectly through the data. This concerns antecedent distance and the overall frequency of referents. The data for these can also be extracted from annotations, but only with a separate R script. For the analysis of antecedent distance, I relied on a previous script written by Nils Schiborr (University of Bamberg), which he graciously shared with me, and in which I only adapted minor details.[17] Concerning the overall frequency of referents, I wrote my own script with the help of Jan H. Boockmann (University of Bamberg).[18] Note that all antecedent-related variables can per default only be used when texts without RefIND are excluded, e.g. *kent02*. In addition, this means that new mentions of referents are excluded in the analysis, since they do not have antecedent distance and would thus distort the results. The languages included in the calculation

[17]The (adapted) script can be found in the Appendix. I am indebted to him for sharing his script with me.

[18]The script can be found in the Appendix.

of the decision trees are Mandarin, Cypriot Greek, Sanzhi, Teop and Vera'a.

In GRAID, zero arguments are annotated using the symbol '0'. Strict rules apply to what counts as a zero argument: it needs to be required by the predicate, an overt referential form in place of zero must be grammatical in the clause, and it needs to be referential (Haig & Schnell 2014: 10). Since many languages, most notably English, have constructions in which an argument is dropped obligatorily, these slots do not qualify as a zero argument, but still have been annotated in GRAID, using 'f0'. As the decision tree is concerned with the CHOICE between referential forms, it would not make any sense to include 'f0' in the analysis. I have thus excluded it.[19]

Note also that there are many different types of pronouns and noun phrases that could make a difference in analysis. For instance, demonstratives are annotated 'dem_pro', possessive pronouns are 'poss_pro' and relative pronouns are 'rel_pro'. Proper names are analysed as noun phrases ('pn_np'). For the sake of simplicity and since this would further reduce the amount of data in each category, I have decided to ignore these smaller distinctions in this thesis, but it might be fruitful for future, more detailed studies with more data to include these distinctions and see if they make a difference. As already discussed in Section 3.2.3.4, I will also include zero arguments annotated as 'svc_0' and 'dt_0' in my analysis.

Regarding the syntactic functions of referents, I slightly adapt the classifications in order to simplify analysis. Predicates ('pred') and vocatives ('voc') are changed to 'other'. Goals ('g') and locatives ('l') are classified as obliques ('obl'). Secondary objects are classified as objects

[19] 'f0' has been included in the barplots in the previous section.

('p'). I exclude all functions with the gloss 'other', 'poss' and 's_ds'. This means that following syntactic functions remain in the analysis: Transitive subject ('a'), intransitive subject ('s'), object ('p'), oblique ('obl'), and secondary object ('p2'). I have excluded dislocated topics in the analysis, since they are never zero arguments.

In the following section, I will now turn to the results of my analysis. First, I will explore the rates of zero arguments in Mandarin and in the other languages, and then complete decision trees for all languages.

4 | Results

In this section, I will apply the quantitative methods explained in the previous section to my data. In the first part, I will investigate the frequency of zero arguments in Mandarin and compare these results to the frequency of zero arguments in the other languages in the corpus. I will thereby consider two perspectives: the distribution of all possible referential forms (noun phrase, pronoun and zero), and the distribution of only pronoun and zero arguments.

In the second part, I turn to the referential choice between noun phrase, pronoun and zero on the one hand, and pronoun and zero on the other hand. For this, I will produce decision trees for Mandarin and, in a second step, for all languages in the corpus to then compare the results.

4.1 Frequency of zero arguments

To assess the frequency of zero arguments in Mandarin in comparison to the other languages in the corpus, I will start with the distribution of noun phrase, pronoun and zero, and then turn to the binary distinction between pronoun and zero. Only subjects, objects, and obliques[20] were included in the analysis, since these are the ones allowing zero arguments in Mandarin.

[20]Including referential forms glossed as goals in GRAID.

Table 2: Referential choice in Multi-CAST

Corpus	*All*	*Zero*	*Pro*	*NP*	*Zero (%)*	*Pro (%)*	*NP (%)*
NKurd	1440	620	119	701	43.06%	8.26%	48.68%
Sanzhi	1137	473	119	545	41.60%	10.47%	47.93%
CypGreek	1216	474	204	538	38.98%	16.78%	44.24%
Mandarin	1596	589	172	835	36.90%	10.78%	52.32%
Tondano	1722	577	455	690	33.51%	26.42%	40.07%
Teop	1724	436	694	594	25.29%	40.26%	34.45%
Veraa	5125	1087	2322	1716	21.21%	45.31%	33.48%
English	2594	170	1190	1234	6.55%	45.88%	47.57%

4.1.1 Distribution of noun phrase, pronoun and zero

In order to answer the first research question,[21] I imported the data from Multi-CAST (Haig & Schnell 2019) using MulticastR (Schiborr 2019b). I then counted all noun phrases, pronouns and zeros. I excluded all instances of these that were glossed *nc* (not classifiable) and all instances of pronouns and zeros in the first and second person, since the noun phrases can only be in the third person. The raw numbers are shown in Table 2. Figure 3 shows the frequency of zero per language in percentages.

As can be seen here, Mandarin is in no respect different from the other languages, and it definitely does not exhibit a higher rate of covertly realised arguments than the other languages. In fact, Sanzhi, Northern Kurdish and Cypriot Greek have a higher rate of zero arguments, Tondano behaves very similarly to Mandarin, and only Teop, Vera'a and English contain a lower rate of zero arguments. In fact, only English has less than 10% of zero arguments, which suggests that earlier research on Mandarin only came to the conclusion that zero arguments were extraordinarily frequent in Mandarin since the language of comparison was English.

[21] Is there a higher rate of zero arguments in Mandarin than in other languages?

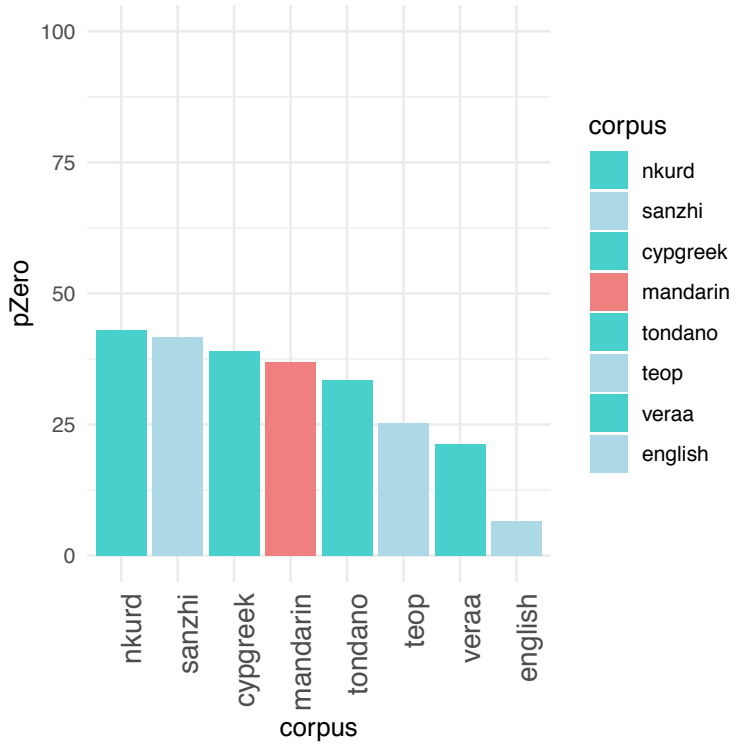

Figure 3: Occurrence of zero in the different languages (%).

If we turn to the percentages of pronouns (Figure 4), we find that the number of pronouns in the corpus is highest for Vera'a, English and Teop, while it is the lowest for Mandarin, Northern Kurdish and Sanzhi, thus exhibiting the opposite of what was shown for percentages of zeros. Interestingly, only around ten percent of all mentions are pronouns in these languages, while zeros make up around 25%.

Finally, regarding lexical noun phrases (Figure 5), we find that differences between languages are comparatively small, with Mandarin showing the highest rate of lexical noun phrases. This cannot be due to 'bad' data in Mandarin, since there is almost no variation between speakers, as Figure 6 shows.

These numbers are supported quite nicely by Pu (1997: 287), who analysed three contemporary novels in order to investigate the distri-

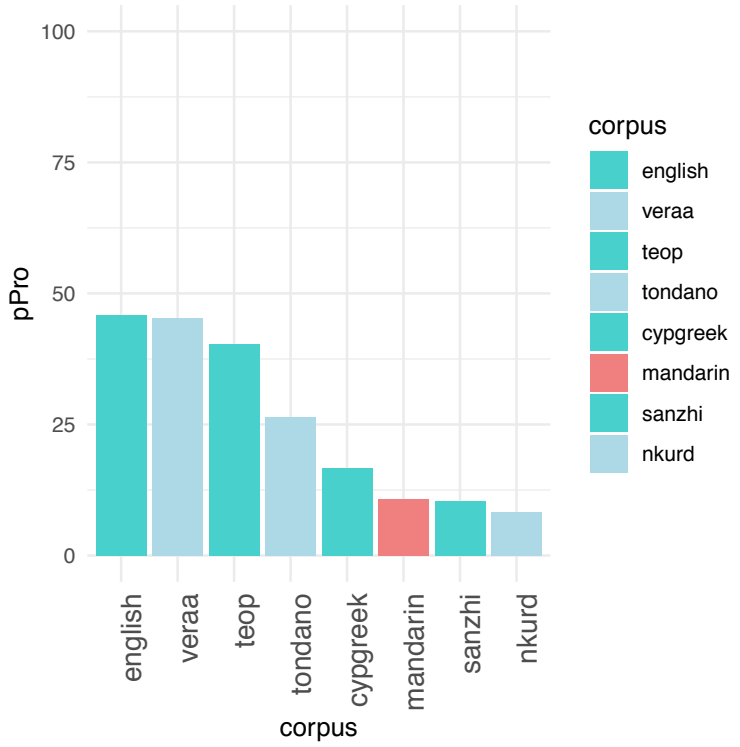

Figure 4: Occurrence of pronouns in the different languages (%).

bution of noun phrases, pronouns and zero arguments. The Mandarin Multi-CAST corpus has 36.9% zero arguments which corresponds to 27%, 26% and 24% in Pu (1997: 287), respectively. The percentage of pronouns in Multi-CAST Mandarin is 11% corresponding to 18%, 11% and 16% respectively. And finally, lexical noun phrases make up 52.32% in Multi-CAST Mandarin and 55%, 61% and 59% in Pu (1997: 287). Thus even in older written literary texts, we still find a similar distribution.

4.1.2 Distribution in different syntactic functions

Since it has often been claimed that Mandarin behaves differently with regard to zero arguments, especially in syntactic functions other than subjects (Battistella 1985: 324, Roberts & Holmberg 2009: 9, Huang

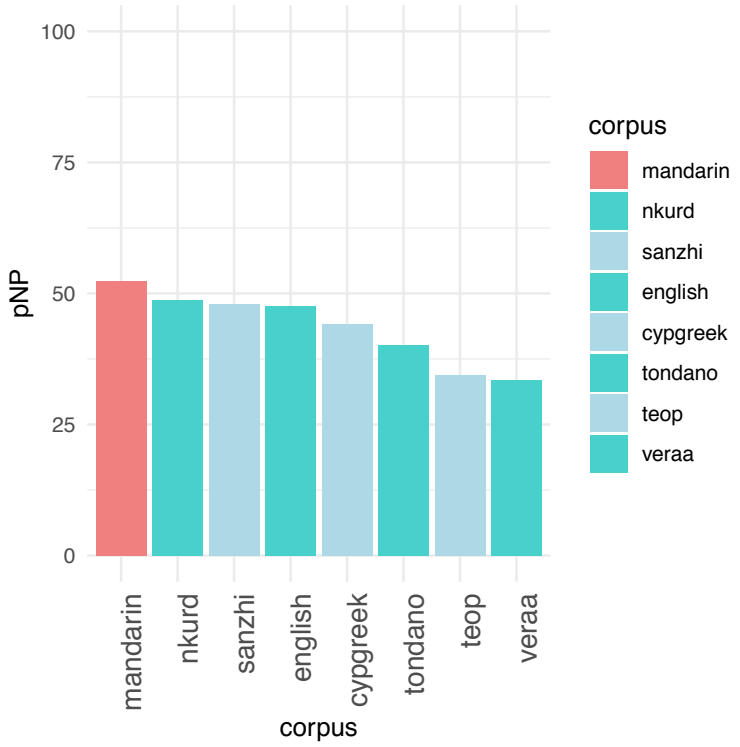

Figure 5: Occurrence of noun phrases in the different languages (%).

1984: 533, Neeleman & Szendröi 2007: 672, Liu 2014), I then divided the data into two subsets: one for subjects,[22] and one for objects and obliques.[23] Note that GRAID (Haig & Schnell 2014) makes a distinction between transitive subjects (A) and intransitive subjects (S), which plays a role in ergative languages and Preferred Argument Structure. However, since this distinction is not relevant in radical pro-drop literature and I did not find significant differences between S and A with regard to the research question, I will sum up these two concepts as subjects for convenience. Figure 7 shows the results for percentages of zeros in the subsample that contains only subjects. As can be seen, the order of languages shows roughly the same pattern as in the full sample (Figure

[22]Excluding all 'nc' and direct speech subjects.
[23]Including referential forms glossed as goals in GRAID.

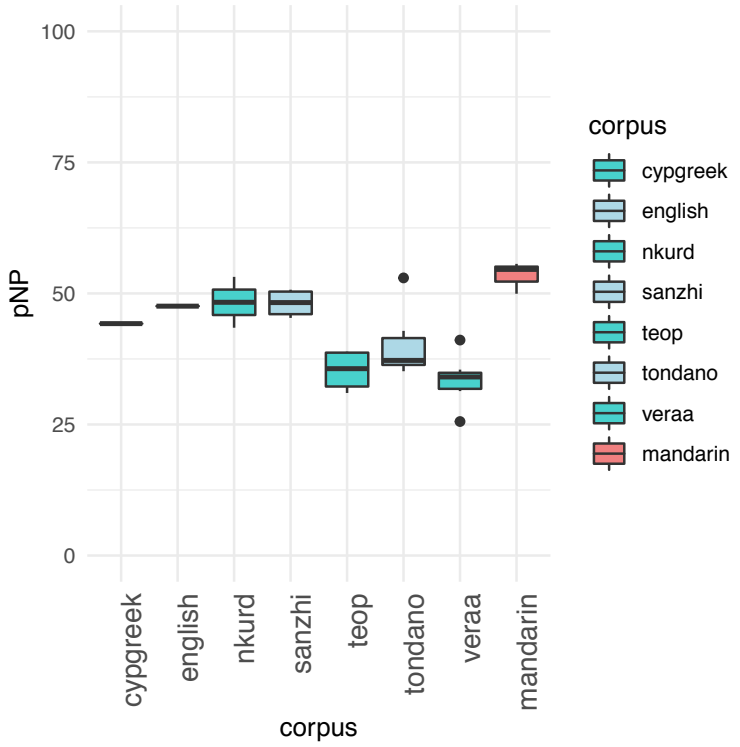

Figure 6: Speaker variation in the production of noun phrases.

3), with Cypriot Greek now exhibiting the highest rate of zero subjects and Tondano now showing a lower number than before. Here, too, Mandarin behaves similarly to the other languages and does not exhibit an extraordinary rate of zero arguments.

Comparing the percentages of pronouns in subject position (Figure 8), we find that English still exhibits the highest percentage of pronouns, and Cypriot Greek the lowest. Mandarin behaves roughly the same as in the full sample. With regard to noun phrases, Figure 9 shows that in all languages subjects show a low number of lexical noun phrases compared to the other functions (objects, goals, obliques) (Figure 5). Mandarin still has the highest percentage compared to the other languages.

In the data subset corresponding to all syntactic functions, I included objects, goals, and obliques, since other functions do not permit zero

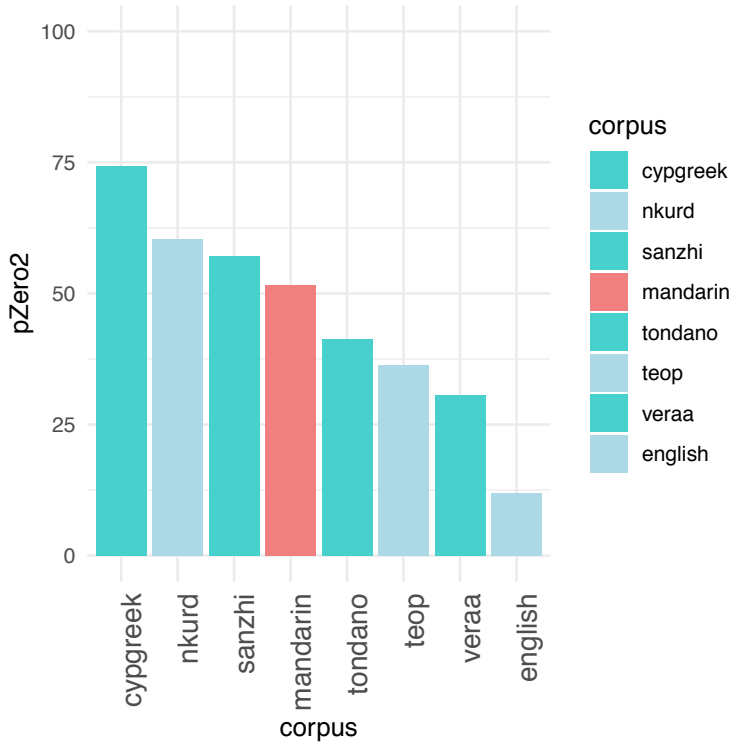

Figure 7: Percentages of zeros in subject position.

arguments. (25) is an example for zero goals:

(25) ## *huibing* *jiu* *gei* *ZERO* *chu* *zhuyi*
bake.pancake ADV ADP 0_Guangliang out idea
np.h:a other adp 0.h:g v:pred np:p
"Pancake-cook put out an idea to (him)." (mandarin_jgz_0450)

Figure 10 shows the percentages of zeros in these functions. As can be seen, the overall numbers of zero arguments in non-subject positions are low in all languages. Compare this to the numbers for only objects (Figure 11), where the overall order of languages remains the same, but with higher numbers of zero arguments.

We can thus conclude that neither in the subject function, nor in the object function or any other syntactic function does Mandarin exhibit an extraordinary rate of zero arguments. The first results and frequency

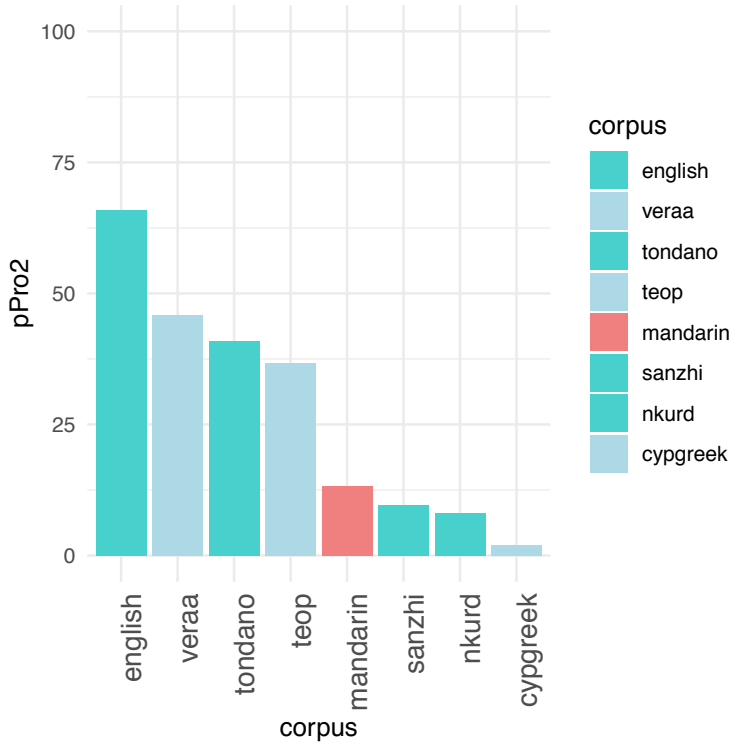

Figure 8: Percentages of pronouns in subject position.

comparisons across eight languages give reason to doubt the claim that Mandarin is special in this regard. If any claim about differences between languages can be made from these first results, tables and figures, then that English appears to stand out with regard to zero arguments. English has the lowest rate of zero arguments, which strengthens the hypothesis that the claim of a particular high number of zero arguments in Mandarin might be due to the English bias of earlier studies.

Interestingly, the frequencies of zero subjects presented in this thesis differ from the numbers from Pu (1997: 287): while over 50% of core arguments in the subject position are zero in my Multi-CAST corpus, this is only the case for 40%, 43% and 38% in Pu (1997: 287). They differ even more with regard to the object: while in the Mandarin corpus the percentage of zeros is almost 25%, it is only 7% and 5% in Pu (1997:

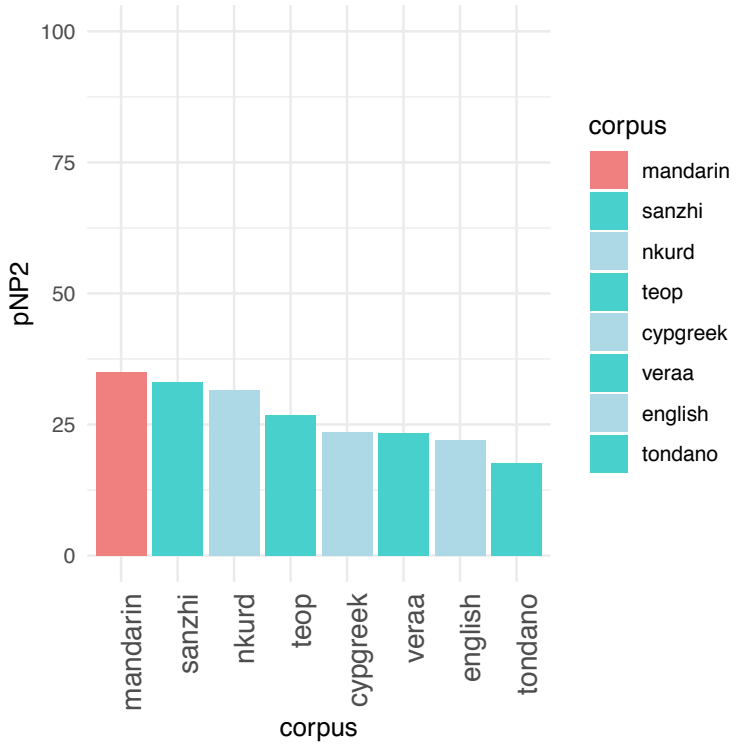

Figure 9: Percentages of noun phrases in subject position.

287). This could be due to the fact that Pu (1997) analysed written novels, while I analyse comparatively natural spoken speech.

4.1.3 Frequency of only pronoun and zero

It could be expected that the difference between languages with regard to the rate of dropped arguments may be higher if lexical noun phrases are excluded. I thus also checked the numbers for only pronouns and zeros, excluding noun phrases. Figure 12 shows a strikingly high difference between Teop, Vera'a and English on the one hand, and Sanzhi, Cypriot Greek, Mandarin, Northern Kurdish and Tondano on the other hand. In Sanzhi, Cypriot Greek, and Mandarin, zero arguments make up almost 75%.

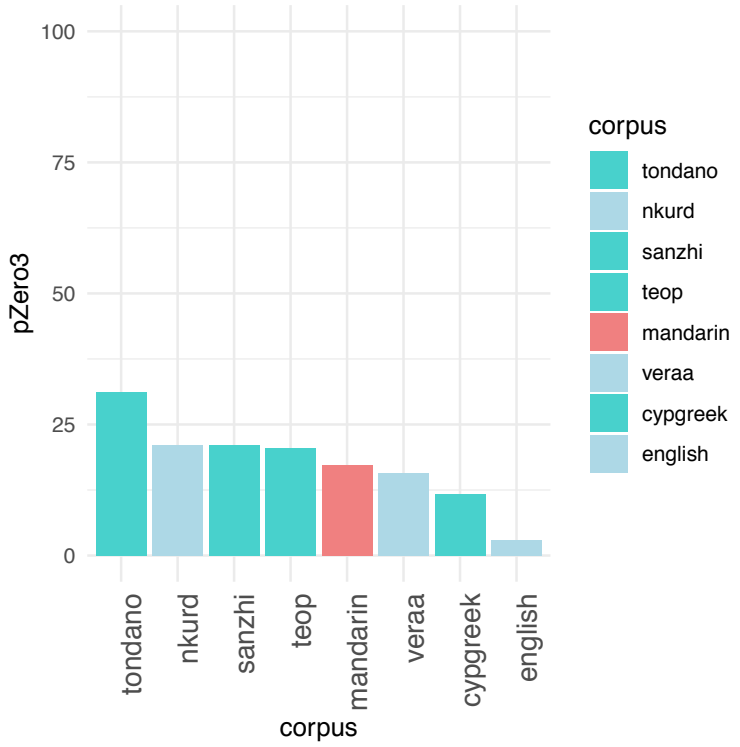

Figure 10: Percentages of zeros in all functions except for subject.

4.1.4 Interim discussion and conclusion

In this section, I have tried to analyse my corpus with regard to my first research question by testing my hypothesis that Mandarin does not have a higher rate of zero arguments than other languages in the corpus.

In order to test this, I first analysed the distribution of noun phrases, pronouns and zero arguments in seven languages of the Multi-CAST corpus (Haig & Schnell 2019) and compared these results to the corpus I had created for Mandarin. I excluded first and second person as well as elements annotated as '*nc*' (non-classifiable) from my data. I found that out of all referential forms in Mandarin, zero arguments make up 37%, which is lower than the numbers for Sanzhi, Northern Kurdish and Cypriot Greek, but higher than the numbers for Tondano, Teop, Vera'a

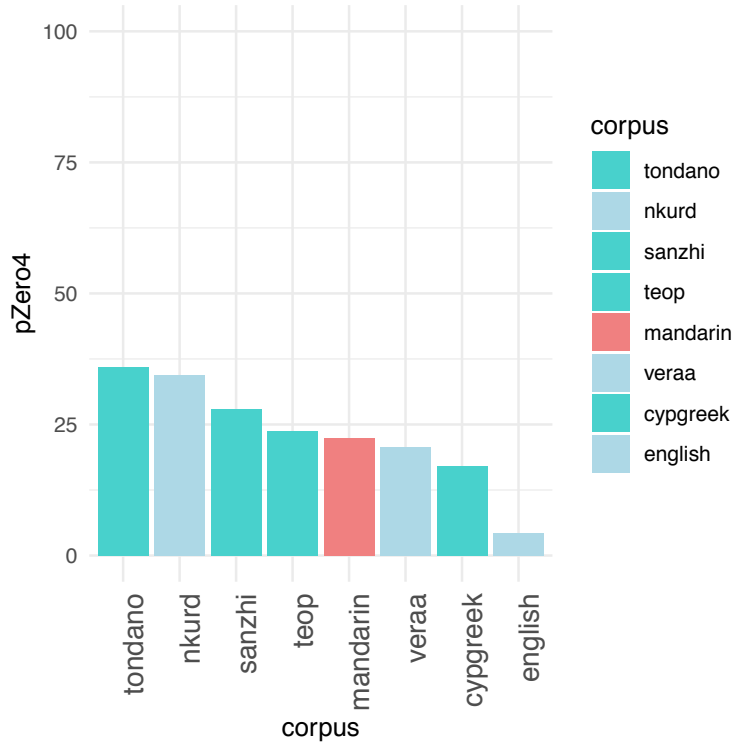

Figure 11: Percentages of zeros in object function.

and English. Except for English, the variance between languages is no higher than ten percent. English only exhibits 6.55% zero arguments. I thus concluded that Mandarin does not show a notably higher rate of zero arguments than other languages in the corpus except for English. Since most previous studies that suggest a special status for Mandarin used English as a baseline, I propose that this might be the reason for believing that Mandarin acts differently.

With regard to pronouns, I found that languages with relatively low numbers of zero arguments had comparatively higher percentages of pronouns instead. Regarding lexical noun phrases however, all languages have similar percentages, albeit with Mandarin exhibiting a slightly higher rate of lexical noun phrases. The reason for this remains unclear for now. It cannot be due to inter-speaker variation since it is very low in Man-

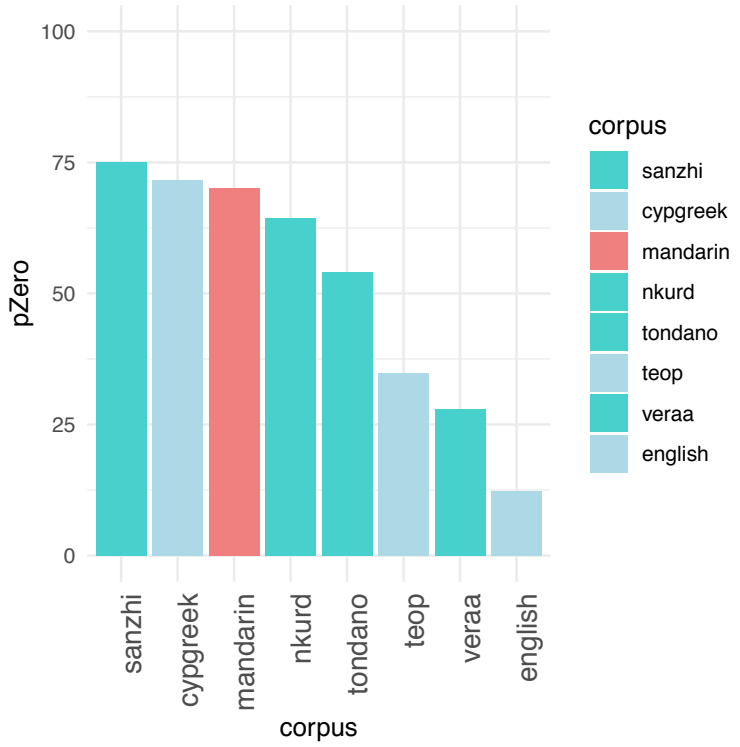

Figure 12: Percentages of zeros in comparison with pronouns.

darin (Figure 6) but also note that the three speakers in my corpus are sociolinguistically-speaking very similar to each other. Another reason might be that proper names might be used like pronouns in Mandarin in some cases but this will have to be tested by future studies. The frequencies of zeros, pronouns and noun phrases in Mandarin are in line with Pu (1997: 287), who counted similar numbers in her study of three contemporary Chinese novels.

Since it has often been claimed in the literature that the difference between pro-drop languages and radical pro-drop languages lies in the omission of arguments in other syntactic functions than the subject, I then turned to the percentages of zeros in different syntactic functions.

With regard to the syntactic function of subject, I found that Mandarin still behaves similarly to the other languages in the corpus, by no

means exhibiting a drastically higher rate of zero arguments. However, I did find a difference between English, Vera'a, Teop and Tondano on the one hand, and Mandarin, Sanzhi, and Northern Kurdish on the other hand. The latter languages show a lower percentage of pronouns in the subject position than the former ones.

I then investigated only the data subset with other syntactic functions (objects, goals, and obliques). I again found that Mandarin behaves very similarly to the other languages in the corpus, some with lower and some with higher percentages of zeros.

When I limited the sample to objects, I even found that while Mandarin stays below 25%, Tondano, Northern Kurdish and Sanzhi have a much higher rate of zero arguments. Once again, only English had a very low number of zero objects. Note that my data currently includes 'f0' (forced zero arguments), thus even omitted arguments that are grammatically enforced.[24] This makes my results even more remarkable.

In the end, I excluded noun phrases from my analysis, and included first and second person. In this way, I investigated the binary distinction between pronoun and zero argument. While Mandarin has the third highest number of zero arguments in the corpus, it is still very much in line with the other languages. I found that Sanzhi, Mandarin, Cypriot Greek and Northern Kurdish exhibit almost 75% of zero arguments, which means that the majority of arguments are zeros in contrast to pronouns. Teop, Vera'a and English reveal percentages that are below the 50% threshold, which means that the majority of arguments in these languages are pronouns in contrast to zero arguments.

Turning back to the initial research question and hypothesis, I conclude that, according to these first results and compared to the eight

[24]They will be excluded for the decision trees, however, since the trees are concerned with referential *choice* and these zeros are obligatory.

languages in the sample, Mandarin does not have a higher rate of zero arguments than other languages in the corpus. With regard to the distinction between pronouns and zeros, it can be said, however, that zero arguments are far more frequent than pronouns in Mandarin discourse. Yet this is not a unique feature of Mandarin alone, and includes Sanzhi, Cypriot Greek, and Northern Kurdish.

4.2 Probabilistic constraints

In Chapter 3.1, I posed the following research question: Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages? I then formulated the hypothesis that probabilistic constraints influence referential choice, and that these constraints are ultimately the same in every language in the corpus. I include following variables and hypotheses for the purpose of this thesis: *Language* and *syntactic function* do not influence referential choice. *Animacy (+/-*HUM*), person, antecedent distance* and the *overall frequency of a referent* influence referential choice.

I use white-box learning, specifically a decision tree using the *rpart* package (Therneau & Atkinson 2019) and the *rpart.plot* package (Milborrow 2019) in R, to find out which of these variables affect referential choice, and if these are the same in every language.

4.2.1 Mandarin

Figure 13 shows the decision tree for Mandarin variation between noun phrase, pronoun and zero. In each box, the middle row informs us of the frequencies of noun phrase, pronoun and zero in the respective data subset. The left number corresponds to the frequency of zero, the middle

number corresponds to the frequency of noun phrases, and the right number corresponds to the frequency of pronouns. The decision tree goes as deeply as possible, which means that even variables that have a small impact are still visible. Variables on the higher nodes thus have a very strong impact, while variables on the lower nodes do not. Another indicator of how important a variable is, is how much of the data it can account for (lowest number in each box). Interestingly, since pronouns only make up 12% of the data (see highest box, middle row, rightmost number), they fall out of the decision tree. This means that the algorithm was not able to find a subset in which the pronouns would be the majority of referential forms according to the variables I have given it.

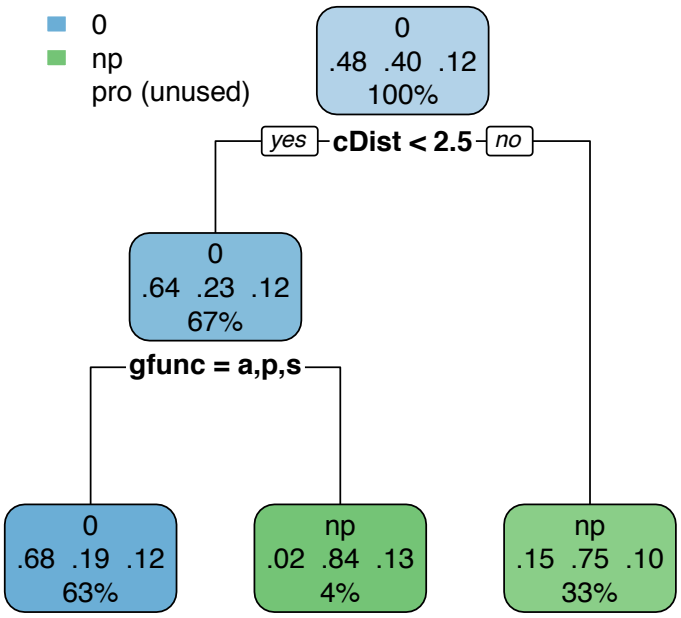

Figure 13: Decision tree for variation between noun phrases and zero arguments in Mandarin. Pronouns are included in the data but unused by the algorithm.

The most important variable in the tree is antecedent distance (= highest node in the tree; cDist). If the last mention of a referent is more than two clauses away (= following the righthand part of the tree, labeled 'no'), the default referential expression is a noun phrase in 75% of all cases (lowest rightmost box in the tree: 'np' gives the information that this is the most frequent referential form, and the middle number in the box tells us the percentage of noun phrases). This classification makes up 33% of the data (lowest number in lowest rightmost box). When the antecedent is mentioned in the previous two clauses (= following the left part of the tree, labeled 'yes'), the default choice is a zero argument with 64% (middle box: '0' tells us that it is the most frequent referential form in this data subset; the frequency of which is told by the leftmost number; the number in the middle (23%) corresponds to the frequency of noun phrases, and the right number (12%) corresponds to pronouns); however, it depends on the syntactic function: if the referent is a subject or object, it will be realised as a zero argument in 68% of all cases, but if it is oblique, it is a noun phrase in 84% of all cases. Note that in all three final nodes, there are still 12%, 13% and 10% of pronouns (= leftmost numbers in each of the lowest boxes, respectively) that cannot be predicted correctly, and similar numbers hold for the noun phrases and zero arguments. There are thus still variables missing that would make a difference in the data. Future studies should include more variables to improve on the decision tree and be able to predict referential forms more accurately.

To sum up, for the decision between noun phrase, zero argument and pronoun, the following variables impact referential choice in Mandarin:

1. Antecedent distance (2.5 clauses)

2. Syntactic function (core arguments vs. adjuncts)

These variables do not impact referential choice:

1. Animacy
2. Overall frequency of referents

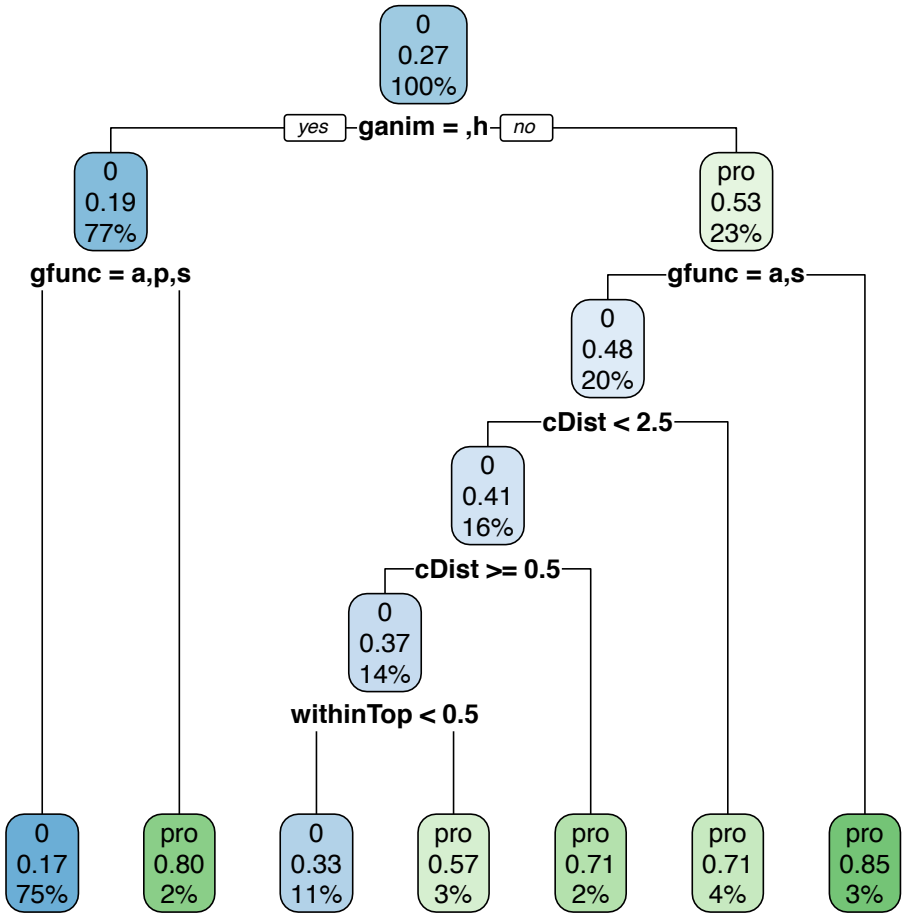

Figure 14: Decision tree for variation between pronouns and zero arguments in Mandarin.

Since the variation between pronoun and zero argument could not be seen directly in the decision tree, I now exclude noun phrases in the

next step, and include first and second person referents in the data. The decision tree is shown in Figure 14.[25] The maximum depth of the tree was reached here again, thus showing all impacting variables, even the ones with minimal impact.

The first distinction in the tree depends on person: If the referent is in the first or second person, it will be in the form of a pronoun in 53% of all cases. It will be a zero argument if the following conditions apply: if it is a subject, if its antecedent is less than three clauses away, and if it is not one of the overall most frequent referents. Note, however, that especially with regard to the last variable, the overall most frequent referents, it is questionable whether this variable should be included. While the other variables correctly predict the majority of pronouns (71%, 71% and 85%, respectively), the overall frequency of referents can only predict 57% of pronouns correctly. Note also that it only appears in the tree if exactly the two most frequent referents are included, but it does not if only the most frequent referent or more than two most frequent referents are included.

If the referent is in the third person, either human or not, subjects and objects are zero arguments by default (83%) and obliques are pronouns (80%).

In conclusion, with regard to variation only between pronouns and zero arguments, the following variables have an impact:

1. Person (third versus first/second)
2. Syntactic function (core versus adjunct, and subject versus object)
3. Antecedent distance (0.5 clauses, 2.5 clauses)

[25] The middle number in each box corresponds to the rate of pronouns in each data subset.

4. (Overall frequency of referent, two most frequent referents in each story, but impact not strong)

Animacy does not have an impact. Note that animacy and person correlate with each other (first and second person is always human in the data) and can thus not be completely separated in the analysis.

Animacy does not play a role in either tree. The overall frequency of referents does not play a role in the first one, and does not have a strong impact in the second one; it thus seems to be relatively irrelevant for both. I subsequently compare these decision trees with the decision trees for all the languages in Multi-CAST (Figure 15).

4.2.2 All languages

This decision tree again includes all variables which have an impact. The variable with the greatest impact is antecedent distance. If a referent's antecedent is more than three clauses away, it will be a noun phrase in 65% of all cases. If it is less than three clauses away, the referential choice depends on the language. In Vera'a and Teop, a human referent will be a pronoun, while an inanimate subject or object will be a zero argument, and an inanimate oblique a noun phrase.

In Cypriot Greek, Sanzhi and Mandarin, a subject will be a zero argument in the majority of cases (74%). If the referent is an object, it will be a noun phrase in Sanzhi and Mandarin (48%) and a pronoun in Cypriot Greek (62%). Note that the two lowest branches, namely (1) the distinction between Sanzhi/Mandarin and Cypriot Greek and (2) the distinction between subject and other functions have a comparatively low impact, since they only account for 7% and 8% of the data, respectively, and can only predict a small majority of cases correctly. To conclude, the following variables seem to play a role in this decision tree:

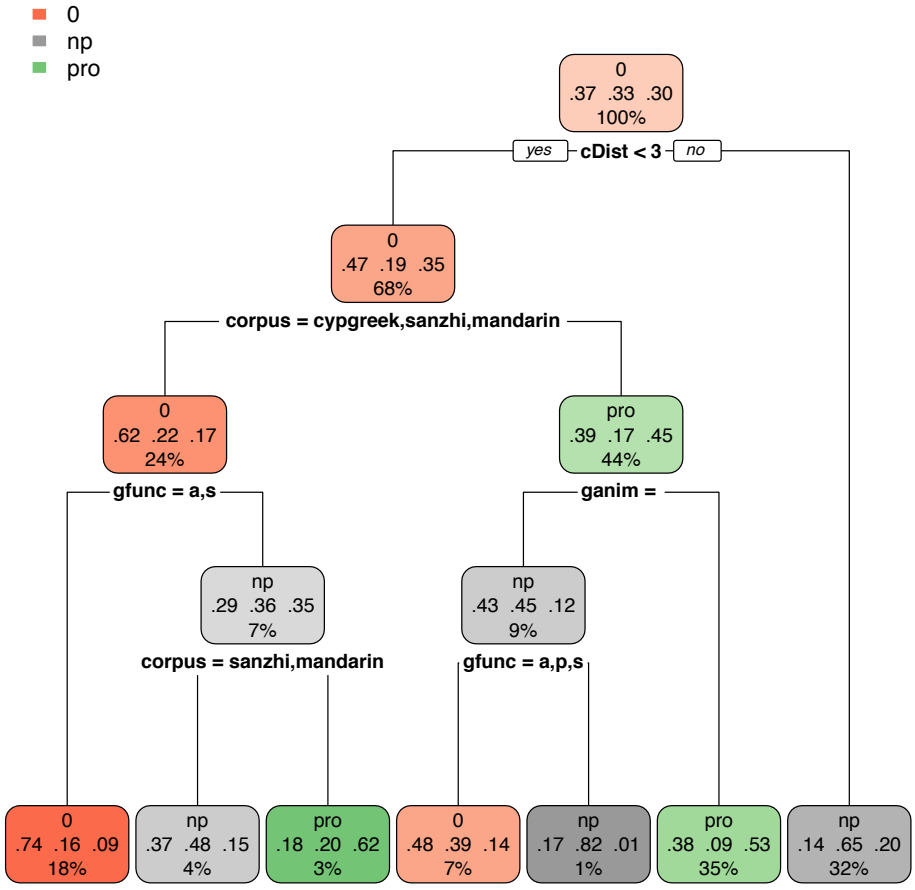

Figure 15: Decision tree for referential choice in all languages.

1. Antecedent distance (three clauses)

2. Language (Mandarin, Cypriot Greek and Sanzhi versus Teop and Vera'a)

3. Syntactic function (subject relevant for Mandarin, Cypriot Greek and Sanzhi; core versus adjunct for the other languages)

4. Animacy (+/-HUM)

The only variable not playing a role is the *overall frequency of referents*. In contrast to my proposed hypothesis, language does play a role in referential choice. While the left part of the tree that is concerned with Cypriot Greek, Sanzhi and Mandarin is only impacted with regard to syntactic function, animacy plays a role in the right part of the tree for the other languages. This is in line with my results for referential choice in Mandarin (Figure 13), where only antecedent distance and syntactic function played a role. We can, however, see that the languages only differ with regard to animacy. All other variables have an impact in all languages, but the way these variables influence referential choice still differs (i.e. syntactic function plays a role in all languages, but in some the distinction lies between core arguments and adjuncts, while it lies in subjecthood for others).

We also find that Mandarin is not the only language that acts differently, but that languages cluster together in groups regarding their behaviour. These clusters conform to the clusters I observed previously in the barplots, showing a tendency for Cypriot Greek, Sanzhi and Mandarin to act similarly.

Lastly, I analysed the variables that impact the decision between pronoun and zero with a decision tree for all languages (Figure 16). In this case, the most important variable is not antecedent distance but rather languages. Again, Cypriot Greek, Sanzhi and Mandarin cluster together. Regarding Vera'a and Teop, the major variable is animacy. If the referent is human, it is a pronoun in 68% of all cases, and if it is not human, it is a zero argument in 79% of all cases. This is the opposite of what I would have expected (i.e. human referents being more topical and thus less marked). Note also that this in case of inanimate referents, only 68% of referents are pronominal and this pertains to 54% of all of my

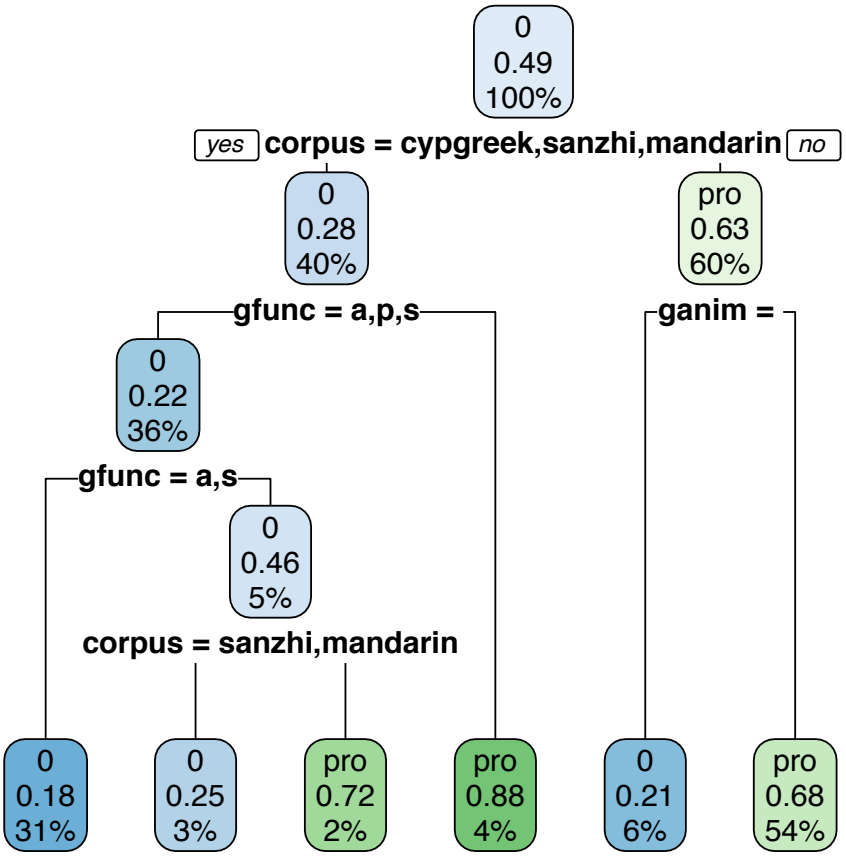

Figure 16: Decision tree between pronoun and zero for all languages.

data. This means that the model wrongly predicts one third of referents in more than half the data. There are thus still variables missing that could improve the prediction of the referential outcome, and the model should be improved in the future.

For Cypriot Greek, Sanzhi and Mandarin, obliques tend to be pronominal (88%). Subjects tend to be zero arguments (82%) and objects are zero arguments as well in Sanzhi and Mandarin (75%), while they are pronominal in Cypriot Greek (72%).

Thus in Sanzhi, Cypriot Greek and Mandarin, syntactic function has

the largest impact on referential choice, and animacy has the most impact in Teop and Vera'a. The languages differ with regard to the impact of animacy. Person, antecedent distance and the overall frequency of referents do not play a role.

The results are similar to the results obtained for noun phrases, pronouns and zero arguments. However, antecedent distance does not play a role now, and the distinction between languages is even more important. This makes sense when looking back to the barplots (Figure 12) where we find that it is the distinction between zero arguments and pronouns where languages differ the most. We find the same cluster of three languages again.

4.2.3 Interim conclusion

This section aimed at answering the research question of which probabilistic constraints influence referential choice and if these constraints are different from constraints in other languages. I hypothesised that probabilistic constraints influence anaphoric distribution and that these constraints are the same in every language. I included the following variables in my analysis: language, syntactic function, humanness, person, antecedent distance and the overall frequency of a referent.

I used decision trees and RStudio (RStudio Team 2018) to find, which variables impact referential choice in which languages. I found that in Mandarin, the choice between noun phrase, pronoun and zero argument is influenced by antecedent distance and syntactic function. If noun phrases are excluded, person, syntactic function and antecedent distance play a role.

Turning to all languages in the corpus, I found that antecedent distance and syntactic function have an impact in all languages. However,

animacy (+/-HUM) does not play a role in Mandarin, Cypriot Greek and Sanzhi. This corresponds to the decision tree on only Mandarin, since animacy (+/-HUM) does not play a role there either. For the distinction between pronouns and zero arguments, Vera'a and Teop make their choice depending on animacy (+/- HUM), while in Sanzhi, Cypriot Greek and Mandarin, syntactic functions are the major variable in deciding referential form.

We can thus conclude that the following variables play a role in referential choice in Mandarin:

1. Antecedent distance
2. Syntactic function
3. Person, if noun phrases are excluded

Animacy (+/-HUM) does not play a role in Mandarin, but plays a role in Teop and Vera'a. These results show that while the constraints are similar in all languages, there are in fact differences between languages. However, Mandarin is not exceptional in this regard but forms a cluster with other languages, which are the same languages that behaved similarly for the analysis of the first research question.[26]

In the next chapter, I will now discuss my results in detail and evaluate these results with regard to their meaning and for previous and future studies.

[26]Is there a higher rate of zero arguments in Mandarin than in other languages?

5 | Discussion of results

At the beginning of the previous chapter, I posed the following two research questions:

1. Is there a higher rate of zero arguments in Mandarin than in other languages?
2. Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages?

   To answer this question, I analysed the following variables:

   a) Languages
   b) Syntactic function
   c) Person
   d) Topicality
      i. Animacy
      ii. Antecedent distance
      iii. Overall frequency of referents

In this section, I will answer the research questions and discuss which variables impact referential choice in which ways.

Is there a higher rate of zero arguments in Mandarin than in other languages? There is a persistent claim that zero arguments in Mandarin are more frequent than in other languages (Li & Thompson 1979, Huang 2000, Yang et al. 2003: 287, Bickel 2003: 708). However, I hypothesised that speakers of Mandarin do not use a higher rate of zero arguments than speakers of most other languages in the corpus. This hypothesis is supported by the results I found on the basis of the corpus data. Mandarin acts similarly to the other languages in the corpus, except for English. In fact, Sanzhi, Northern Kurdish and Cypriot Greek exhibit a higher percentage of zero arguments (Figure 3).

Regarding radical pro-drop in the sense that Mandarin allows pro-drop in all syntactic functions (Battistella 1985: 324, Roberts & Holmberg 2009: 9, Huang 1984: 533, Neeleman & Szendröi 2007: 672, Liu 2014), the raw numbers show that Mandarin does exhibit zero arguments for subjects, objects and obliques[27]. Yet, this is true for other languages as well, and, as in other languages, zero arguments are most frequent for subjects. However, given the choice between pronoun and zero arguments, Sanzhi, Mandarin, Cypriot Greek and Kurdish speakers choose zero arguments in the majority of cases, while Tondano, Teop, Vera'a and English speakers choose pronouns in the majority of cases. In this sense, then, the data supoort the claim that zero arguments are the default referential choice in Mandarin rather than pronouns (supporting Li & Thompson 1979, Pu 1997: 281 and Battistella 1985: 324).

Which probabilistic constraints influence referential choice? In Section 3, I hypothesised that probabilistic constraints influence referential choice. Since the decision trees can correctly predict the majority of the data even though I had to exclude some variables due to lack of

[27]Including referential forms glossed as goals in GRAID.

space and time, this hypothesis proves to be true. With regard to the variables, I suggested that they would be the same in every language in the corpus. However, they differ among languages. I also proposed that syntactic function would not influence referential choice. However, even though it does not impact it in the way Du Bois (2003) proposed, the distinction between subject, object and other non-core arguments has an impact. With regard to topicality (= animacy, antecedent distance and overall frequency of referents) and person, I suggested that they influence referential choice. Indeed, person and antecedent distance impact referential choice in some cases, while animacy only impacts some languages. Overall frequency of referents does not appear to be a decisive factor.

In Mandarin, syntactic function and antecedent distance impact referential choice. If the choice is only between pronoun and zero, an additional constraint is person.

The cutoff point for the decision between referential forms with regard to antecedent distance is 2 clauses. If a referent has previously been mentioned in the last 2 clauses, it tends to be a zero argument, while it is a noun phrase if the last mention is more distant. This thus supports claims of accessibility theory, which suggests that more topical referents are less marked (2.3.1.5). Interestingly, antecedent distance also plays a role in the distinction between zero and pronoun for inanimate referents in Mandarin, contrary to findings by Schnell & Barth (2018: 76).[28] With regard to syntactic function, the core arguments subject and object form a unit contrasting with obliques and goals, and further down the tree, subjects and objects behave differently.

I will now discuss in detail which variables have an impact on referential choice according to my results, and which variables do not have

[28]However, this is only true for Mandarin. Antecedent distance is not relevant for the choice between pronoun and zero when looking at all languages (Figure 16.

an impact. A variable proposed to impact referential choice is **animacy** (Fraurud 1996, Ariel 1996: 22, Hsiao et al. 2014). Pu (1997: 290) finds that animacy increases the likelihood of pronouns in contrast to zero arguments in both English and Mandarin. Li (2012: 102) also notes that animacy impacts referential choice for Mandarin pronominal subjects. Schnell & Barth (2018) note that the effect of animacy might be due to discourse topicality rather than animacy itself. In my findings, however, animacy does not impact referential choice in Mandarin, Sanzhi and Cypriot Greek, while it does affect referential choice in Vera'a and Teop, in line with what Schnell & Barth (2018) found for Vera'a. Unfortunately, I do not know why there is a difference between languages and hope that future studies will be able to shed more light on this. A potential reason could be that animacy does in fact indirectly express topicality, and that its effect was thus more strongly captured in one of the other variables connected to topicality.

Topicality has often been claimed to influence referential choice (Ariel 1996: 22, Schnell & Barth 2018: 73, Huang 1984: 541, Ackema et al. 2006: 15). I proposed that, since topicality is hard to define in a quantitatively measurable way, it is connected to three variables in my study: animacy, antecedent distance and overall frequency of referents. With regard to **antecedent distance**, then, I found that it does play a role in referential choice in Mandarin. However, topicality in the sense of discourse topicality as discussed in Schnell & Barth (2018: 59) would have been connected to animacy and this variable did not prove significant in the Mandarin data.

Overall frequency of referents also did not have a decisive impact in my study. Since I proposed three variables that could possibly express topicality (animacy, antecedent distance, and overall frequency

of referents), and since antecedent distance had the greatest impact, it is possible that the effect of topicality was simply already captured in antecedent distance and thus did not show again at a different point in Mandarin, Cypriot Greek, and Sanzhi.

Several studies (Wratil 2011, Ariel 1996, Li & Bayley 2018) have noted that **person** might be connected to referential choice. Wratil (2011: 119) and Ariel (1996: 22) note that person is connected to topicality in that first person is more topical than second person, and second person is more topical than third person. The most topical referent would most likely be the least marked, which would be zero in Mandarin. Li (2012: 102) and Li & Bayley (2018: 149) showed that person plays a role in subject omissions in Mandarin. My data supports this finding, since person impacts the choice between pronoun and zero. However, I found that, contrary to Wratil (2011), first and second person tend to be pronominal, while third person tends to be zero. This might be due to different registers, however, since in my narratives the first and second person pronouns do not refer to the speech participants (the speaker and his/her audience), but to the protagonists in the story.

Syntactic function might play a role in referential choice. Most famously, Du Bois (1987: 823, see also 2003: 34) claims that languages show an ergative pattern in their distribution of new and lexical arguments. They avoid lexical transitive subjects, but prefer lexical intransitive subjects and objects. Haig & Schnell (2016) show that these patterns are due to animacy, however. While I do find that syntactic function impacts referential choice in all languages, transitive and intransitive subjects behave similarly and have to be seen in contrast with objects and oblique arguments, which thus does not support Du Bois (1987, 2003).

Are these constraints different from constraints in other languages? The question if all languages exhibit the same constraints in referential choice has been posed several times, and opinions vary. For instance, it has been noted that rich or poor verbal inflection might change the way hearers retrieve referents, and thus change the way speakers choose referential forms (Ackema & Neeleman 2007, Ackema et al. 2006: 15).

Yet, previous studies have suggested that constraints are the same across languages: Pu (1995) showed that the same constraints can explain referential choice in English and Mandarin, and Torres Cacoullos & Travis (2019) showed the same for English and Spanish.

However, in my study, I found that while all languages behave similarly with regard to some constraints (i.e. syntactic function, antecedent distance), they do not behave similarly with regard to all (i.e. animacy).

The same or very similar constraints hold for Cypriot Greek, Mandarin and Sanzhi on the one hand, and Teop and Vera'a on the other hand. In light of the results presented in Torres Cacoullos & Travis (2019: 682), it would be interesting to add English and Spanish to the corpus in the future, so that they could be compared to Mandarin.

In conclusion, while all languages exhibit pragmatic constraints, there are clear differences between languages, and they seem to cluster together in certain ways. However, further research is needed to test and support these findings. Specifically, more data is needed to substantiate these findings, with more variables that could change the decision tree and predict outcomes more correctly, and better statistical methods, e.g. the use of training and testing data subsets and random forests.

Since Sanzhi, Mandarin and Cypriot Greek behave similarly with regard to the decision trees, the question is if they also act similarly with

respect to the barplots and raw numbers of percentages. Going back to the first research question,[29] we find that they cluster together in the barplots as well. Their percentages are especially close with regard to their distribution of pronoun and zero (Figure 12) and the distribution of referential forms of objects and subjects (e.g. Figure 8). Kurdish clusters with Mandarin, Cypriot Greek and Sanzhi in almost all cases, but since it did not have RefIND at the time of analysis, it is excluded in the decision trees. It was, however, published with RefIND shortly before the thesis was submitted and preliminary analysis points in the direction that it does cluster with Mandarin, Sanzhi and Cypriot Greek, as would be expected.

Unfortunately, RefIND was not available for English at the time of analysis and writing of the thesis, and thus had to be excluded from the decision trees. Since the raw numbers and percentages showed that English differs from Mandarin the most, it would be very interesting to compare these two languages with regard to probabilistic constraints. I hope that the necessary data for this will be made available soon and expect interesting results in the comparison. Nils Schiborr has kindly granted me access to preliminary data from the English sub-corpus to be published in the 1908 version of Multi-CAST. A preliminary calculation of the decision tree with only Mandarin and English as data points yields antecedent distance as the most important variable, and the next node in the tree differentiates between English and Mandarin. This is what we expect if English clusters with Teop and Vera'a (see Figure 15 for comparison) and it at least very preliminarily suggests that my line of analysis and interpretation could hold true for more languages, including English.

[29] Is there a higher rate of zero arguments in Mandarin than in other languages?

Thus Huang's (2000: 261-277) point for a typology of "pragmatic languages" might be worth pursuing. While of course all languages are pragmatic, I do find clusters of languages that behave in certain patterns, but the reason why these patterns come into being and if this argumentation persists even after further studies can only be explored by future research.

Turning back to Huang's (1992: 27) claim that zero arguments in Mandarin are pragmatically determined, while in English they are grammatically determined, we find that all languages in the corpus exhibit probabilistic and pragmatic constraints that impact referential choice. However, we also find that these constraints differ between languages, and that their scope in allowing zero arguments greatly differs cross-linguistically.

6 | Conclusion and outlook

The title of this thesis poses the question of how radical pro-drop in Mandarin really is. In order to answer this question, I quantitatively analysed various languages in Multi-CAST (Haig & Schnell 2019) regarding zero arguments and referential choice, with a special focus on Mandarin.

I have provided a detailed discussion of the theoretical background, most notably of definitions of pro-drop and radical pro-drop and of which factors influence referential choice according to recent studies. I then posed two research questions, namely *1) Is there a higher rate of zero arguments in Mandarin than in other languages?*, and *2) Which probabilistic constraints influence referential choice, and are these constraints different from constraints in other languages?* Comparing the frequencies of zero arguments in the Mandarin sub-corpus to the other sub-corpora, I found that Mandarin does not differ significantly from other languages, and that there are some languages that exhibit zero arguments more frequently than Mandarin (contrary to Li & Thompson 1979, Huang 2000).

I did find, however, that Mandarin does exhibit zero anaphora in all syntactic functions, and that compared to other languages like English, it belongs to a group of languages that prefer zero arguments over pronouns in the majority of cases (supporting e.g. Battistella 1985: 324, Pu 1997: 281).

With regard to probabilistic constraints, I used decision trees that

make it possible to include more than one variable and to account for correlations between them. I found that the two most important variables were syntactic function and antecedent distance in Mandarin. Note that I did not find a difference between transitive and intransitive subjects with regard to syntactic function, thus finding no support for Du Bois (1987) and Du Bois (2003). Person influences the choice between pronoun and zero. Comparing these results to the other sub-corpora, I found that languages behave differently with regard to probabilistic constraints. Most notably, Sanzhi, Mandarin and Cypriot Greek clustered together, while Teop and Vera'a were in addition impacted by animacy.

Thus even though I found that Mandarin behaves similar to other languages and is not "special", I also find that differences between languages exist that might be worth pursuing in future research. Most notably, languages seem to cluster together with regard to their frequency of zero arguments and probabilistic constraints in referential choice. It would be worth including more languages in the corpus, most notably English as the classically studied non-pro-drop language, Kurdish, since it clusters with Mandarin in the barplots, Spanish as the classically studied pro-drop language, as well as some geographically more diverse languages in South America and Africa. More variables should also be included, since I had to limit myself to six (language, syntactic function, animacy, person, antecedent distance, overall frequency of referent). Other interesting variables that could not be included here are for instance discourse segmentation (Giora & Lee 1996: 114), definiteness (Ariel 1996: 22), constructions with specific semantic verb classes (Travis & Cacoullos 2012: 725), TAM (Travis & Cacoullos 2012: 725) and other antecedent-related factors like the antecedent's syntactic function and form. Li (2012: 102) even notes that sociolinguistic factors of the speakers make a difference

but, since my speakers show almost no variation (Figure 6) and are all male university students between 20 and 30 years of age, this could not be analysed in my study. It might be worth considering once more stories with other speakers have been added to Multi-CAST.

Another question that should be answered in future research is if rich or poor verbal inflection influences probabilistic constraints in referential choice, i.e. if languages with rich verbal inflection cluster together with regard to referential choice, and languages with poor verbal inflection cluster together with regard to referential choice (see e.g. Li & Bayley 2018: 137, Ackema & Neeleman 2007, Ackema et al. 2006: 15). Another question is if Huang's (2000: 261-277) call for a typology of "pragmatic languages" makes sense with regard to the clustering of certain languages with each other; especially because all languages showed probabilistic pragmatic constraints, but differed with regard to the specific constraints.

Regarding the statistical analysis, this thesis had some methodological constraints. I had to limit myself to raw numbers, frequencies and decision trees, excluding a conditional random forest analysis and the division of my data into a test and training subset. My results should be tested further and supported by taking a step further into these statistical analyses.

Other questions connected to referential choice and anaphoric resolution could not be included in this thesis, even though they would be very interesting to pursue, especially in light of the new Mandarin subcorpus in Multi-CAST, i.e. the use of the Chinese reflexive *ziji* (see e.g. Battistella 1985 and Huang 2000).

Recently, Chambaz & Desagulier (2016) have called for a more unified approach between statistics and (corpus) linguistics, believing that less boundaries between the disciplines can help tackle problems that have

been extensively discussed in the literature. I hope that this thesis provides a step further in this direction in showing that statistics and corpus linguistics can help us understand actual language use and probabilistic choices made by speakers.

References

Ackema, Peter; Patrick Brandt; Maaike Schoorlemmer; and Fred Weermann. 2006. The role of agreement in the expression of arguments. Ackema, Peter et al. (eds.), *Arguments and agreement,* 1–32. Oxford: OUP.

Ackema, Peter, and Ad Neeleman. 2007. Restricted pro drop in Early Modern Dutch. *The journal of comparative Germanic linguistics* 10. 81–107.

Adams, Marianne. 1987. From Old French to the theory of pro-drop. *Natural language and linguistic theory* 5. 1–32.

Adibifar, Shirin. 2016. Multi-CAST Persian. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Adibifar, Shirin. 2019. Multi-CAST Persian annotation notes. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Ariel, Mira. 1988. Referring and accessibility. *J. Linguistics* 24. 65–87.

Ariel, Mira. 1996. Referring expressions and the +/- coreference distinction. Fretheim, Thorstein and Jeanette K. Gundel (eds.), *Reference*

and referent accessibility, 13–35. Amsterdam/Philadelphia: John Benjamins.

Barbosa, Pilar. 2009. Two kinds of subject *pro*. *Studia linguistica* 63(1). 2–58.

Barbosa, Pilar. 2011a. *Pro*-drop and theories of *pro* in the Minimalist program part 2: Pronoun deletion analyses of null subjects and partial, discourse and semi *pro*-drop. *Language and linguistics compass 5/8.* 571–587.

Barbosa, Pilar P. 2011b. *Pro*-drop and Theories of *pro* in the Minimalist program part 1: Consistent null subject languages and the pronominal-agr hypothesis. *Language and linguistics compass 5/8.* 551–570.

Battistella, Edwin. 1985. On the distribution of pro in Chinese. *Natural language and linguistic theory* 3. 317–340.

Bennis, Hans. 2006. Agreement, pro, and imperatives. Ackema, Peter et al. (eds.), *Arguments and agreement,* 101–123. Oxford: OUP.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Bresnan, Joan; Anna Cueni; Tatiana Nikitina; and R. Harald Baayen. 2005. Predicting the dative alternation. *KNAW Academy Colloquium: Cognitive foundations of interpretation,* October 27-28, 2004, Amsterdam. Corrected September 24, 2005.

Brickell, Timothy. 2016. Multi-CAST Tondano. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

REFERENCES

Brugman, H., and A. Russel. 2004. Annotating Multimedia/ Multimodal resources with ELAN. *Proceedings of LREC 2004, Fourth International Conference on Languages Resources and Evaluation.*

Chafe, Wallace L. (ed.). 1980. *The pear stories. Cognitive, cultural, and linguistic aspects of narrative production.* Norwood, New Jersey: ABLEX (= Advances in discourse processes; 3).

Chambaz, Antoine, and Guillaume Desagulier. 2016. Predicting is not explaining: Targeted learning of the dative alternation. *Journal of causal inference* 4(1). 1–30.

Chomsky, Noam. 1981. *Lectures on Government and Binding.* Dordrecht, Cinnaminson: Foris Publications.

Dahl, Östen, and Kari Fraurud. 1996. Animacy in grammar and discourse. Fretheim, Thorstein and Jeanette K. Gundel (eds.), *Reference and referent accessibility,* 47–64. Amsterdam/Philadelphia: John Benjamins.

Dowle, Matt, and Arun Srinivasan. 2019. *data.table: Extension of 'data.frame'.* R package version 1.12.2, https://CRAN.R-project.org/package=data.table.

Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63(4). 805-855.

Du Bois, John W. 2003. Argument structure: Grammar in use. Du Bois, John W, Lorraine E. Kumpf, and William J. Ashby (eds.), *Preferred argument structure. Grammar as architecture for function*, 11–60. Amsterdam/Philadelphia: John Benjamins (= Studies in Discourse and Grammar; 14).

ELAN Version 5.2 [Computer software]. 2018, April 4. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/.

Forker, Diana, and Nils N. Schiborr. 2019. Multi-CAST Sanzhi Dargwa. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts. (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Fraurud, Kari. 1996. Cognitive ontology and NP form. Fretheim, Thorstein and Jeanette K. Gundel (eds.), *Reference and referent accessibility,* 65–88. Amsterdam/Philadelphia: John Benjamins.

Fuss, Eric. 2011. Historical pathways to null subjects: Implications for the theory of pro-drop. Wratil, Melani and Peter Gallmann (eds.), *Null pronouns,* 53–98. Berlin, Boston: de Gruyter (Studies in Generative Grammar; 106).

Gelormini-Lezama, Carlos. 2018. Exploring the repeated name penalty and the overt pronoun penalty in Spanish. *Journal of psycholinguistic research* 47, 377–389.

Giora, Rachel, and Cher-Leng Lee. 1996. Written discourse segmentation: the function of unstressed pronouns in Mandarin Chinese. Fretheim, Thorstein and Jeanette K. Gundel (eds.), *Reference and referent accessibility,* 113–140. Amsterdam/Philadelphia: John Benjamins.

Hadjidas, Harris, and Maria Vollmer. 2015. Multi-CAST Cypriot Greek. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Haig, Geoffrey. 2018. 2.3. Northern Kurdish (Kurmanjî). Haig, Geoffrey and Geoffrey Khan, *The languages and linguistics of Western Asia,* 106–158. Berlin, Boston: De Gruyter.

Haig, Geoffrey, and Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual Version 7.0. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Haig, Geoffrey, and Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.

Haig, Geoffrey, and Stefan Schnell. 2018 [2016]. Multi-CAST research context. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Haig, Geoffrey, and Stefan Schnell. 2019. *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de) (02.08.2019).*

Haig, Geoffrey; Stefan Schnell; and Nils Schiborr. 2017. The limits of accessibility. Talk held at *ALT12,* ANU Canberra, 7 December 2017.

Haig, Geoffrey; Maria Vollmer; and Hanna Thiele. 2019a. Multi-CAST Northern Kurdish. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Haig, Geoffrey; Maria Vollmer; and Hanna Thiele. 2019b. Multi-CAST Northern Kurdish annotation notes. Haig, Geoffrey and

Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural language & linguistic theory* 1(1). 5-47.

Harrell Jr, Frank E. 2019. *rms: Regression modeling strategies.* R package version 5.1-3, https://CRAN.R-project.org/package=rms.

Hsiao, Yaling; Yannan Gao; and Maryellen C. MacDonald. 2014. Agent-patient similarity affects sentence structure in language production: Evidence from subject omissions in Mandarin. *Frontiers in psychology* 5. 1–12.

Huang, James C.-T. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry* 15(4). 531–574.

Huang, Yan. 1992. Against Chomsky's typology of empty categories. *Journal of pragmatics* 17. 1–29.

Huang, Yan. 2000. *Anaphora. A cross-linguistic study.* Oxford: OUP (= Oxford Studies in Typology and Linguistic Theory).

Iemmolo, Giorgio, and Giorgio Francesco Arcodia. 2014. Differential object marking and identifiability of the referent: A study of Mandarin Chinese. *Linguistics* 52(2). 315–334.

Kibrik, Andrej A. 2011. *Reference in discourse.* Oxford: OUP.

Kimoto, Yukinori. 2018. Operationalizing Philippine-type syntax for the GRAID system: Clause structure, case marking, and verb class in Arta. *Asian and African languages and linguistics* 12. 17–35.

Koeneman, Olaf. 2006. Deriving the difference between full and partial pro-drop. Ackema, Peter et al. (eds.), *Arguments and agreement,* 76–100. Oxford: OUP.

Li, Charles N., and Sandra A. Thompson. 1976. Subject and topic: a new typology of language. Li, Charles N. (ed.), *Subject and topic,* 457–490. New York, San Francisco, London: Academic Press.

Li, Charles N., and Sandra A. Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. Givón, Talmy (ed.), *Discourse and syntax,* 311–336. New York: Academic Press (= Syntax and Semantics; 12).

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese. A functional reference grammar.* Berkeley, Los Angeles, London: University of California Press.

Li, Xiaoshi. 2012. Variation of subject pronominal expression in Mandarin Chinese. *Sociolinguistic studies* 6(1). 91–119.

Li, Xiaoshi, and Robert Bayley. 2018. Lexical frequency and syntactic variation. Subject pronoun use in Mandarin Chinese. *Asia-Pacific language variation* 4(2).

Liu, Chi-Ming Louis. 2014. *A modular theory of radical pro drop.* Doctoral dissertation, Harvard University. Cambridge, Massachusetts: Harvard University.

Liu, Feng-Hsi. 2007. Word order variation and *ba* sentences in Chinese. *Studies in language* 31(3). 649–682.

Milborrow, Stephen. 2019. *rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart'.* R package version 3.0.7, https://CRAN.R-project.org/package=rpart.plot.

Mosel, Ulrike, and Stefan Schnell. 2015. Multi-CAST Teop. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Neeleman, Ad, and Kriszta Szendröi. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry* 38(4). 671–714.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes.* R package version 1.1-2, https://CRAN.R-project.org/package=RColorBrewer.

Ooms, Jeroen. 2019. *curl: A modern and flexible web client for R.* R package version 3.3, https://CRAN.R-project.org/package=curl.

Perlmutter, David. 1971. *Deep and Surface Structure constraints in syntax.* New York: Holt, Rinehart and Winston.

Pu, Ming-Ming. 1995. Anaphoric patterning in English and Mandarin narrative production. *Discourse processes* 19. 279–300.

Pu, Ming-Ming. 1997. Zero anaphora and grammatical relations in Mandarin. Givón, Talmy (ed.), *Grammatical relations: A Functionalist perspective.* John Benjamins.

R Core Team. 2019. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/.

Roberts, Ian, and Anders Holmberg. 2009. Introduction: parameters in Minimalist Theory. Biberauer, Theresa, Anders Holmberg, Ian Roberts and Michelle Sheehan (eds.)*, Parametric variation. Null subjects in Minimalist Theory,* 1–57. Cambridge: CUP.

Rosenkvist, Henrik. 2010. Null referential subjects in Övdalian. *Nordic journal of linguistics* 33(3). 231–267.

RStudio Team. 2018. *RStudio: Integrated development environment for R.* Boston, MA: RStudio, Inc., http://www.rstudio.com/.

Schiborr, Nils N. 2015. Multi-CAST English. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Schiborr, Nils N. 2016. Multi-CAST corpus overview and description. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Schiborr, Nils N. 2019a. Multi-CAST collection overview. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (25. 08. 2019).*

Schiborr, Nils N.; Stefan Schnell; and Hanna Thiele. 2018. RefIND — Referent indexing in natural-language discourse. Annotation guidelines v1.1. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Schiborr, Nils Norman. 2018. Data-driven models of referential choice. Antecedent distance and beyond. *Talk held at ISSLaC3, 7 December 2018.*

Schiborr, Nils Norman. 2019b. multicastR: A companion to the Multi-CAST collection. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts,*

https://CRAN.R-project.org/package=multicastR, R package version 1.1.0.

Schnell, Stefan. 2015. Multi-CAST Vera'a. Haig, Geoffrey and Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts (multicast.aspra.uni-bamberg.de/) (02. 08. 2019).*

Schnell, Stefan, and Danielle Barth. 2018. Discourse motivations for pronominal and zero objects across registers in Vera'a. *Language variation and change* 30. 51–81.

Schnell, Stefan, and Nils N. Schiborr. 2018. Corpus-based typological research in discourse and grammar: GRAID and Multi-CAST. *Asian and African languages and linguistics* 12. 1–16.

Schnell, Stefan; Nils Norman Schiborr; and Geoffrey Haig. 2018. Is intransitive subject the preferred role for introducing new referents? Evidence from corpus-based typology. Talk held at *SLE 2018*, 01 September 2018.

Sessarego, Sandro, and Javier Gutiérrez-Rexach. 2017. Revisiting the Null Subject Parameter: New insights from Afro-Peruvian Spanish. *Isogloss* 3(1). 43–68.

Soares, Eduardo Correa. 2016. Yes-No answers, partial pro-drop languages and machine translation. *Procedia social and behavioral sciences* 231. 135–142.

Speas, Margaret. 2006. Economy, agreement, and the representation of null arguments. Ackema, Peter et al. (eds.), *Arguments and agreement,* 35–75. Oxford: OUP.

Stoll, Sabine, and Balthasar Bickel. 2009. How deep are differences in referential density? Guo, Jiansheng et al. (eds.), *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin,* 543–555. New York, London: Psychology Press.

Sun, Chaofen. 2006. *Chinese. A linguistic introduction.* Cambridge: CUP.

Therneau, Terry, and Beth Atkinson. 2019. *rpart: Recursive partitioning and regression trees.* R package version 4.1-15, https://CRAN.R-project.org/package=rpart.

Torres Cacoullos, Rena, and Catherine E. Travis. 2019. Variationist typology: shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3).

Travis, Catherine E., and Rena Torres Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive linguistics* 23(4). 711–748.

Wang, Longyue; Zhaopeng Tu; Xiaojun Zhang; Siyou Liu; Hang Li; Andy Way; and Qun Liu. 2017. A novel and robust approach for pro-drop language translation. *Mach translat* 31. 65–87.

White, Lydia. 1985. The "pro-drop" parameter in adult second language acquisition. *Language learning* 35(1). 47–62.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York, https://ggplot2.tidyverse.org.

Wratil, Melani. 2011. Uncovered *pro* — On the development and identification of null subjects. Wratil, Melani and Peter Gallmann

(eds.), *Null pronouns,* 99–140. Berlin, Boston: de Gruyter (=Studies in Generative Grammar; 106).

Yang, Chin Lung; Peter C. Gordon; Randall Hendrick; and Chih Wei Hue. 2003. Constraining the comprehension of pronominal expressions in Chinese. *Cognition* 86. 283–315.

Zhang, Ting Liu, Weinan; Yin Qingyu; and Zhang Yu. 2019. Neural recovery machine for Chinese dropped pronoun. *Frontiers of computer science* 13(5). 1023–1033.

Appendix

Due to lack of space and the complexity of my data, please find all additional material (R-scripts, ELAN-files, etc.) used in this thesis on the accompanying CD. It contains the three stories taken from my Mandarin sup-corpus (ELAN-files and WAV-files), as well as the R-scripts. There is one script for the barplots ('R_Script_Barplots.R'), two for the decision trees ('R_Script_Trees_1/2.R'), and two for the extraction of the variables antecedent distance and frequency of referents, respectively. The four tsv-tables containing the raw data (list of referents, metadata, Mandarin raw data, raw data of all other languages). With regard to the tsv-table with the raw data of all other languages, I only included the table on which I based my initial analysis, not including the tables from the preliminary analysis on Kurdish and English. Interested readers are referred to the Multi-CAST website (www.multicast.aspra.uni-bamberg.de, accessed: 26.08.2019) where all necessary additional data can be downloaded.

Erklärung

Ich erkläre hiermit gemäß § 22 Abs. 2 i.V.m. § 19 Abs. 2 APO, dass ich die vorstehende Masterarbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden, dass Zitate kenntlich gemacht sind und die Arbeit noch in keinem anderen Prüfungsverfahren vorgelegt wurde und dass die in unveränderbarer maschinenlesbarer Form eingereichte Fassung mit der schriftlichen Fassung identisch ist.

30. August 2019 Maria Vollmer

(Datum) (Unterschrift)