Investigating Theories of Speaker Choice in a Classifier Language

by

Meilin Zhan

B.A., Guangdong University of Foreign Studies (2012)
M.Sc., The University of Edinburgh (2013)
M.A., University of California, San Diego (2016)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author . . . .
Department of Brain and Cognitive Sciences
August 13, 2020

Certified by . . . .
Roger Levy
Associate Professor
Thesis Supervisor

Accepted by . . . .
Rebecca Saxe
John W. Jarve (1978) Professor of Brain and Cognitive Sciences
Associate Head, Department of Brain and Cognitive Sciences
Affiliate, McGovern Institute for Brain Science

Investigating Theories of Speaker Choice in a Classifier Language

by

Meilin Zhan

Submitted to the Department of Brain and Cognitive Sciences
on August 13, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

Speakers often face choices as to how to structure their intended message into an utterance. When multiple options are available to express more or less the same meaning, what general principles govern speaker choice? Here I investigate the influence of contextual predictability on the encoding of linguistic content manifested by speaker choice in a classifier language, Mandarin Chinese.

In English, a numeral modifies a noun directly (e.g., three tables). In classifier languages such as Mandarin Chinese, it is obligatory to use a classifier (CL) with the numeral and the noun (e.g., three CL.flat table, three CL.general table). While different nouns are compatible with different specific classifiers, there is a general classifier "ge" (CL.general) that can be used with most nouns. I focus on the alternating options between using the general classifier versus a specific classifier with the same noun where the options are nearly semantically invariant.

When the upcoming noun has high surprisal, the use of a specific classifier would reduce surprisal at the noun thus potentially facilitate comprehension (predicted by the Uniform Information Density account (Levy & Jaeger, 2007)), but the use of that specific classifier may be dispreferred from a production standpoint if accessing the general classifier requires less effort (predicted by the Availability-Based Production account (Bock, 1987; Ferreira & Dell, 2000)). Using evidence from naturalistic language datasets and real-time language production experiments, I have argued that speaker choice is emerged from availability-based production and is the results of speaker-centric information processing, where availability is affected by linguistic experience such as contextual predictability and word frequency, as well as real-time production pressure in speaking. From the perspective of language design, the redundancy in information content shared between nouns and classifiers provides a mechanism that potentially is beneficial for the listener, making communication more robust. However, the capacity of using a specific classifier is constrained by speaker-oriented availability production.

Thesis Supervisor: Roger Levy
Title: Associate Professor

Acknowledgments

First and foremost, I want to thank my advisor Roger Levy, one of the most passionate, curious, and admirable scientists I have ever met. It has been a rare privilege to have Roger as an advisor for the past few years. He has always encouraged me to keep perspectives, to seek the most important questions, and to hold myself to the highest standards. And he has done all this in a kind and supportive way.

I am grateful to have a wonderful thesis committee. Thanks are due to my committee chair Ted Gibson, who has provided inspiring suggestions and moral support for the development of this thesis, and for my journey at MIT more generally. Thanks to Josh Tenenbaum for his pinpointed suggestions that have helped me situate my own work in relation to other work in the field. Thanks to Naomi Feldman for her detailed and insightful feedback on this work.

This thesis would not have been possible without the help of these undergraduate research assistants: Jiaxing Liu, Jamie Geng, Siyi Lin, and Silvia Cho.

I want to thank past and present members of MIT Computational Psycholinguistics Lab for their support during my grad school journey: Helena Aparicio, Yevgeni Berzak, Veronica Boyce, Richard Futrell, Jon Gauthier, Matthias Hofer, Jenn Hu, Titus von der Malsburg, Stephan Meylan, Emily Morgan, Till Poppels, Peng Qian, and MH Tessler.

I would also like to thank professors and scientists who have supported, inspired, and mentored me along the way, especially Sarah Brown-Schmidt, Gary Dell, Evelina Fedorenko, Victor Ferreira, Andrew Kehler, and Hannah Rohde.

I am fortunate to have made and kept some dear friends during my time at MIT. I am grateful for their support and friendship, especially: Fengdi Guo, Di Jin, Jinghui Liu, Xiaofan Xu, Guangyi Yue, and Verena Hehl.

Finally, I want to express my deepest gratitude to my parents for their unconditional love and support from thousands of miles away. To my best friend and husband, Wenzhe (a.k.a. Cosmo), who has been a constant source of inspiration in my life: Thank you for your advice and encouragement at all hours and on all fronts.

The research presented in this thesis was funded by NSF research grants BCS-1456081 and BCS-1551866 to Roger Levy, BCS-1844723 to Roger Levy and Meilin Zhan, and an MIT Henry E. Singleton Fellowship to Meilin Zhan.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Natural language is propagated through the interactions of production, comprehension, learning and cultural transmission among generations of language users, and is shaped by cognitive, communicative, and cultural pressures (MacDonald, 2013; Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Gibson et al., 2019; Hawkins, 1994, 2014). In this thesis, I will focus on language production, one of the shaping forces contributing to why language is structured the way it looks today.

Why do speakers say the things they say in the moment? The simple answer is that largely, speakers say the things they say to express their intended meaning. Speaking begins with the formulation of a preverbal MESSAGE of the speaker's communicative intention and ends with the articulation of selected WORD FORMS (W. J. Levelt, 1989). If we consider the speaker as a highly complex information processor, the formulation component translates a conceptual structure to a linguistic structure. During this process, words are retrieved from the lexicon and organized in a linear sequence to be produced (grammatical encoding), and phonetic plans are created for articulation for the sequence (phonology encoding) (W. J. Levelt, 1989). Choice points are available throughout the production process, from conceptualization of the message, formulation of the message including selecting and ordering words, to the articulation of the selected word forms. An average speaker produces around 150 words per minute at a normal speech rate, and this number can be doubled when speaking under time pressure (Maclay & Osgood, 1959). In a cognitive process like speaking, the decision rate is high. Still, the error rate is low, about 1 slip of tongue per 1,000 words in spontaneous conversations under normal condition (Garnham, Shillcock, Brown, Mill, & Cutler, 1981). How does the production system manage to process such a large amount of information and select intended linguistic expressions effectively?

Depending on the communicative goal, speakers can choose different words or organize words in various ways to convey their intended messages. Furthermore, meaning is not the only factor influencing what words they choose to say or not to say. The expressivity of natural language often provides multiple options for speakers to convey more or less the same meaning. For example, sentences (1)–(3) express the same or nearly the same meaning. The difference between (1)/(2) and (3) is in the choice of content words, and the difference between (1) and (2) is in the choice of syntactic structure and function words.

(1) Sheldon wrote a thank-you letter to Penny.

(2) Sheldon wrote Penny a thank-you letter.

(3) Sheldon wrote Penny a letter of appreciation.

As shown in the above examples, speakers have options to choose what words or labels to use, as well as how to organize those words into phrases and sentences. This is what I consider as SPEAKER CHOICE in language production in this thesis. (3) and (4) exemplify another type of optionality in language, OPTIONAL REDUCTION in part of a sentence.

(4) The research assistant *that* the director hired did a great job.

(5) The research assistant the director hired did a great job.

In the case of optional reduction above, speakers can choose to produce the function word *that* or to omit it without changing the meaning of the utterance. This is another example of speaker choice.

Statement of the problem When multiple options are available to express more or less the same meaning, what general principles govern a speaker's choice? There are twin pressures of comprehender-oriented optimization (H. H. Clark & Brennan, 1991; S. E. Brennan & Clark, 1996; Galati & Brennan, 2010; Levy & Jaeger, 2007, sometimes stated with reference to AUDIENCE DESIGN), and production-based speaker-oriented optimization (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gahl, 2008; Gahl, Yao, & Johnson, 2012; V. S. Ferreira & Dell, 2000) operating on speakers. To what extent do speakers choose options that potentially facilitate comprehension, and to what extent does speaker choice emerges as a by-product of speaker-oriented information processing?

Studies have shown that both speakers and listeners are sensitive to probability distribution in language. For comprehension, probabilistic predictability effects are ubiquitous in disambiguation (Jurafsky, 1996; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; MacDonald, Pearlmutter, & Seidenberg, 1994) and modulating incremental processing difficulty (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013; J. R. Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016a; Shain, Blank, van Schijndel, Schuler, & Fedorenko, 2020; Staub & Clifton Jr, 2006). For production, speakers adjust their speech with regard to probability distribution in language (J. Bybee, 2002, 2003; Dell & Gordon, 2003; Gahl & Garnsey, 2004; Levy & Jaeger, 2007; Jaeger, 2010; Gahl et al., 2012). Probabilistic models for language, specifically models for contextual predictability, has paved a promising path to investigating what general principles govern speaker choice.

One of the earliest observations of efficient use of language shows that word length and frequency are inversely correlated: frequent words are generally shorter (Zipf, 1949). The idea behind this relationship is that speakers use short words more often than long ones because short words require least effort to produce. More recent studies have shown that optional reduction in language is sensitive to contextual predictability: for many linguistic phenomena, greater contextual predictability yields a greater rate of reduction. Predictable words tend to be shorter in length (Piantadosi, Tily, & Gibson, 2011), and the same word is more likely to appear in a reduced form (if a reduced version exists for that word, e.g., *mathematics* vs. *math*) when it is predictable from the context (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Predictable words, either predictable in the local context or relatively predictable due to overall high frequency, are produced with shorter duration and with less phonetic details such as consonant lenition and vowel centralization (Lieberman, 1963; Jurafsky, Bell, Gregory, & Raymond, 2001; Aylett & Turk, 2004; Gahl & Garnsey, 2004; Gahl, 2008; Bell et al., 2009; Seyfarth, 2014). For example, Lieberman (1963) observed that the tokens of the word *nine* had a shorter duration and were less intelligible when produced in a highly predictable context (e.g., *A stitch in time saves...*) than a less predictable context (e.g., *The next word will be ...*).

At the morphosyntactic level, predictability effects have been observed in optional SYNTACTIC REDUCTION with similar trends: higher predictability yields higher rate of reduction (Levy & Jaeger, 2007; Jaeger, 2010; Wasow, Jaeger, & Orr, 2011; V. S. Ferreira & Dell, 2000; A. F. Frank & Jaeger, 2008; Barth, 2019). Observed syntactic reduction phenomena include but are not limited to English auxiliary contraction (6) (A. F. Frank & Jaeger, 2008; Barth, 2019), infinite marker omission (7) (Lohmann, 2011), relativizer *that* omission (8) (Levy & Jaeger, 2007; Wasow et al., 2011), complementizer *that* omission (V. S. Ferreira & Dell, 2000; Jaeger, 2010) (9), and optional object case-marking in Japanese (Kurumada & Jaeger, 2015).

(6) {She is / She's} coming to the wedding tomorrow.

(7) Peter helped us (to) feed our dog when we were traveling.

(8) The family (that) you work for is very wealthy.

(9) Jane believes (that) she will get a promotion soon.

There are different theories accounting for speaker's preference in syntactic variation, such as dependency length minimization (Gibson, 1998; Temperley, 2007; Futrell, Mahowald, & Gibson, 2015; Futrell, Gibson, & Levy, 2020), syntactic priming

(J. K. Bock, 1986; Pickering & Ferreira, 2008), availability effects (K. Bock, 1987; V. S. Ferreira & Dell, 2000), and uniform information density (Levy & Jaeger, 2007; Jaeger, 2010). Yet it is unclear whether predictability-based optional reduction is geared toward enhancing comprehension, or toward facilitating production.

For optional function word omission such as (8) and (9), predictability effects have been argued to be consistent with both a speaker-oriented account (V. S. Ferreira & Dell, 2000; K. Bock, 1987), where speakers use *that* to buy in more time especially when the upcoming phrase is less available (also less predictable on the assumption that less predictable materials are also harder or take longer to become available), and a potentially listener-oriented account (Levy & Jaeger, 2007; Jaeger, 2010), where speakers use *that* to offload some information from an otherwise less predictable onset of the upcoming phrase to avoid overwhelming the comprehender. These accounts have proven difficult to disentangle empirically.

Proposed solution in this thesis Here I investigate theories of speaker choice by looking at speaker choice in a classifier language, Mandarin Chinese, where it is obligatory to use a classifier (CL) with the numeral and the noun (e.g., three CL.flat table, three CL.general table). While different nouns are compatible with different specific classifiers, there is a general classifier *ge* (CL.general) that can be used with most nouns. I focus on the alternating options between using the general classifier versus a specific classifier with the same noun where the options are nearly semantically invariant. When the upcoming noun has high surprisal, the use of a specific classifier would reduce surprisal at the noun thus potentially facilitate comprehension (predicted by the Uniform Information Density account (Levy & Jaeger, 2007)), but the use of that specific classifier may be less preferred from a production standpoint if accessing the general classifier requires less effort (predicted by the Availability-Based Production account (V. S. Ferreira & Dell, 2000; K. Bock, 1987)).

Using a combination of corpus study and behavioral experiments, I found that the distribution of speaker choices supports the speaker-oriented account. These results offer converging evidence regarding the mechanisms influencing speaker choice, and help advance our understanding of the cognitive processes involved in converting thoughts to speech.

1.1 Predictability effects on language production

Predictability influences language comprehension and production Most of the time, language carries communicative value and is produced to be understood. Contextual predictability plays an important role in both language comprehension and production: Predictable materials would facilitate comprehension, and may affect the way speakers produce utterances. Previous work has shown that language users are highly sensitive to contextual predictability (see Kuperber and Jaeger (2016) for a review). Reaction times to predictable words are faster compared to unpredictable words in a variety of behavioral tasks in the domains of lexical decision (Fischler & Bloom, 1979; Forster, 1981; Stanovich & West, 1983; Schwanenflugel & LaCount, 1988), phrasal deicison (Arnon & Snider, 2010), naming (Forster, 1981; McClelland & O'Regan, 1981; Traxler & Foss, 2000), and speech monitoring (Cole & Perfetti, 1980; Marslen-Wilson, Brown, & Tyler, 1988). Readers fixate less on predictable words than unpredictable ones (Balota, Pollatsek, & Rayner, 1985; Rayner, Binder, Ashby, & Pollatsek, 2001; Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; S. L. Frank & Bod, 2011; Smith & Levy, 2013). There is also evidence from neurolinguistics indicates that language processing mechanisms in the brain generate predictions about upcoming words in naturalistic sentence comprehension (J. R. Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016b; Shain et al., 2020).

Speakers might aim to produce words in a less predictable context with longer duration and higher intelligibility (Lieberman, 1963; Bell et al., 2003; Aylett & Turk, 2004; Bell et al., 2009; Seyfarth, 2014). And this is likely to benefit comprehension. However, speakers have limited cognitive resources and incomplete knowledge of the world and of the interlocutor's mental state. When speakers are planning for complex utterances, they tend to produce fillers such as *uh* and *um* (H. H. Clark & Fox Tree, 2002), repeat themselves (H. H. Clark & Wasow, 1998), produce intonational phrase boundaries (D. Watson & Gibson, 2004), and lengthen the duration of the words (Bell et al., 2003, 2009; D. G. Watson, Arnold, & Tanenhaus, 2008). At the syntactic level, optional function words such as *that* are more likely to be produced when the phrase they introduce is less available or predictable (V. S. Ferreira & Dell, 2000; Levy & Jaeger, 2007; Jaeger, 2010; Wasow et al., 2011). These results indicate that when speech is effortful, the duration of the words tend to be greater. More generally, words and sentences tend to appear in their full linguistic form.

Taking results from both language comprehension and production into account, we can see the following trend: High predictability generally makes words easy to comprehend and easy to say in a shortened form at the same time.

Operationalization of predictability As predictability effects serve as a vehicle to connect language comprehension and production, investigating what role predictability effects play in language production and how they shape speaker choice in principled ways can shed light on the central principles guiding language production. First, how can one operationalize the notion of predictability in language?

Here I adopt a broad notion of predictability, including how predictable the words are in the local context and if context is minimal, how predictable the words are in general based on their frequencies in daily language use. For naturalistic data, I use SURPRISAL, defined as the negative log probability of the word in the context (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; S. L. Frank & Bod, 2011; Smith & Levy, 2013) (Equation 1.1) estimated using language models. If words are completely predictable from the context, then the surprisal would be zero (zero bit of information gained). Conversely, if words are less predictable from the context, they will have a high surprisal value. Here I use surprisal as the linking function to estimate contextual predictability of a given linguistic unit.

$$\textsc{Surprisal} = -\log_2 P(W_i|\textsc{Context}) \tag{1.1}$$

For out-of-context picture-naming experiments, I use word frequency from `SogouW` (Sogou, 2006), a word frequency dictionary for online texts in Chinese.

For optional function word omission such as (10) , predictability effects have been argued to be consistent with both the speaker-oriented account of Availability-Based Production (K. Bock, 1987; V. S. Ferreira & Dell, 2000), and the potentially listener-oriented account of Uniform Information Density (Levy & Jaeger, 2007; Jaeger, 2010).

(10) James designed a dress (that) jazz dancers like.

Both accounts predict that the less predictable the clause introduced by the function word, the more likely the speaker should be to produce the function word *that*. Predictability here refers to how hard it is to predict the target word from the context, measured by information content contained by the target word. Given the context, if the target word contains high information content, then it is more surprising and harder to predict. This notion of predictability is most naturally associated with a comprehender's perspective: if a word is less predictable, it means that it is harder for the comprehender to guess what it is (before hearing or seeing it) from the context.

However, predictability effects are also relevant to the speaker and affect speaker behavior. When the listener had a hard time predicting the next word given the context, the original speakers also tended to experience production difficulty and are more likely to hesitate at those points (Goldman-Eisler, 1958). Using a Tic Tac Toe game, researchers found that disfluencies and intonational boundaries, which are linked to production difficulty, were more likely to occur in less predictable moves than in predictable moves (D. G. Watson et al., 2008). These results indicate there is a link between predictability and availability of a word: the less predictable a word from the context, the less available it is when it comes to producing that word. Consequently, the speaker is likely to encounter production difficulty such as disfluencies and longer response latencies. For out-of-context picture-naming, the time it takes to name an object is highly correlated with the frequency of its name occurring in the language: the response latency for lower frequency nouns are longer than high frequency nouns (Oldfield & Wingfield, 1965). Based on the above-mentioned research, it is safe to say that language users have some shared linguistic experience and they are sensitive to probability distribution in language. What is easy to comprehend is often easy to say and vice versa.

Below I illustrate how the two accounts predict speaker choices with regard to optional reduction using the above example (10).

1.1.1 Speaker-oriented account: Availability-Based Production

Availability-Based Production (ABP) proposes that production is more efficient if speakers choose options that allow quickly selected lemmas to be mentioned as soon as possible (K. Bock, 1987; V. S. Ferreira & Dell, 2000). Lemmas here refer to abstract conceptual representations of to-be-produced words in early stages of language production, a stage between a message and realized word forms in grammatical encoding (W. J. Levelt, 1989; V. S. Ferreira & Dell, 2000; Warren, 2013). It is argued that the timing of lemma selection affects syntactic structures. If the lemma that encodes a part of the message is selected and becomes available at the time to convey that part of the message, the corresponding word form will be produced. If that lemma is not yet available, then other available lemmas will be used, as long as it is compatible with what the speaker has said thus far and it does not interfere with the speaker's future path to expressing the desired content. This is also referred to as **the principle of immediate mention** (V. S. Ferreira & Dell, 2000). This principle is aligned with the notion of **incremental production**, where speakers plan as they speak and the production of an utterance is piece by piece (W. J. Levelt, 1989; V. S. Ferreira, 1996; Brown-Schmidt & Tanenhaus, 2006; Brown-Schmidt & Konopka, 2015).

Suppose a speaker has selected the word *dress* in (10) and is in the process of conveying the remainder of the utterance meaning as a relative clause (RC). If the word *jazz* becomes available quickly, the function word *that* will likely be omitted. If the word *jazz* does not become available quickly, the speaker will produce the function word *that* to buy more time for *jazz* to become available, assuming the word *that* generally becomes available quickly as it is a high-frequency word used in a wide variety of contexts. Previous research has observed a relationship between contextual predictability and speech production: fluent speech consisted of sequences of words which were easily predicted from the context by guessers who knew little about the original speaker's intentions with regard to these utterances, however, where guessers had a hard time predict the next word given the context, the original speakers also tended to hesitate at those points (Goldman-Eisler, 1958). Taking this relationship into account, I assume that less predictable words are harder or take longer to become available in production. This assumption is also consistent with previous work. It is known that lemma accessibility is driven by noun frequency from picture naming (Oldfield & Wingfield, 1965), and even if the word is given (for the speakers to read), lower frequency words still take longer to produce (Balota & Chumbley, 1985). Based on this assumption, the less predictable the relative clause, the lower the probability that its first word *jazz* will be available at the time when the speaker is about to produce the RC. Thus the speaker will be more likely to use *that*. Here an RC is required after *dress* in order for it to be followed by the word *jazz*. So the lower the contextual probability of an RC, the lower the contextual probability of its first word, predicting the observed relationship between phrasal onset probability and optional function word omission rate.

In short, Availability-Based Production predicts that optional function word *that* will be used more often with less predictable relative clauses than predictable relative clauses.

1.1.2 Listener-oriented account: Uniform Information Density

Uniform Information Density (UID) proposes that within boundaries defined by grammar, when multiple options are available to encode the message, speakers prefer the variant that distributes information density more uniformly (lower variance), to decrease the chance of information loss or miscommunication (Levy & Jaeger, 2007; Jaeger, 2010). Multiple formalizations are possible under this account (Genzel & Charniak, 2002; Aylett & Turk, 2004; Maurits, Navarro, & Perfors, 2010; Levy, 2018). Here I will adopt the definition of the UID hypothesis defined in Levy

(2018). In this version of UID, there are two components in the cost function: 1) peaks and troughs in information density are penalized; 2) long strings are penalized. For $k >= 1$ and $c >= 0$, if an utterance consists of a string of symbol sequence $w_{1 \ldots n}$, the cost function is defined as Equation 1.2:

$$C(w_{1 \ldots n}) = \sum_{i=1}^{n} [-logP(w_i|w_{1 \ldots i-1})]^k + cn \tag{1.2}$$

where as the value of k increases, the more heavily non-uniformity of information density through the string is penalized; and increasing the value of c penalizes longer strings. [1]

In (10), if the function word *that* is omitted, the first word of the relative clause (*jazz* in this example) is highly unpredictable and would convey two pieces of information: both the onset of the RC and part of the content of the RC itself. These both contribute to the information content of the word *jazz*, which can be measured using surprisal (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013). If the function word *that* is produced, having *that* at the onset of the RC splits these two pieces of information apart, offloading the RC's onset onto *that* and now *jazz* only conveys RC-internal content. In this way, the word *jazz* would have lower information content, potentially avoiding a peak in information density and thus facilitate comprehension.

In summary, Uniform Information Density predicts that the optional function word *that* will be used more often with less predictable relative clauses than predictable relative clauses.

As is shown above, both the Availability-Based Production and the Uniform Information Density make the same prediction in the case of optional reduction of the function word *that* in English. The two accounts are not easily distinguishable based on the data observed in *that*-omission.

[1] Note that although we are being precise and providing the mathematical formulation of the UID account, the predictions that we make are qualitative.

1.2 Roadmap

The remainder of this thesis is organized as follows.

- Chapter 2 introduces the classifier system in Mandarin Chinese, defines the scope of classifiers in this study, provides a brief synchronic overview of classifier languages and a diachronic overview of the evolution of classifier, and illustrates predictions of the two accounts for Mandarin classifiers.

- Chapter 3 presents a corpus study investigating the relationship between contextual predictability and classifier choice using naturalistic data. This work was published as Zhan and Levy (2018). The dissertation author was the primary investigator and author of this paper.

- Chapter 4 presents two controlled experiments investigating classifier choice in real-time production by manipulating noun frequency and time pressure imposed on the speaker in settings with and without a human listener present. Part of this work is based on Zhan and Levy (2019). The dissertation author was the primary investigator and author of this paper.

- Chapter 5 presents a picture-naming experiment investigating potential effects of counting on classifier production by manipulating the number of objects shown in the picture.

- Chapter 6 summarizes the thesis and provides directions for future work.

Chapter 2

Chinese Classifiers as a Test Case for Theories of Language Production

2.1 Numeral classifiers

2.1.1 The scope of classifiers in this thesis

In English and other languages, COUNT NOUNS, such as *apple*, and *chair*, generally can be used with numerals directly to form phrases such as *two apples*, and *three chairs*. MASS NOUNS, such as *water*, and *sand*, usually do not occur directly with numerals. Instead, additional elements are needed to construct a phrase with numerals, turning to phrases such as *two bottles of water*, and *three cups of sand*. These additional elements are sometimes referred to as **MENSURAL NUMERAL CLASSIFIERS**, as they provide a unit of measure for quantity (Gil, 2013).

In Mandarin and other languages, an additional grammatical element is required even for count nouns. As shown in Examples (1) and (2), in Mandarin Chinese, an additional grammatical element (marked as CL for numeral classifier) is required when a noun occurs with a numeral.

(1) 两 个 苹果
liang ge pingguo
two CL apple ("two apples")

(2) 三 把 椅子
san ba yizi
three CL chair ("three chairs")

Such grammatical elements are often referred to as **SORTAL NUMERAL CLASSIFIERS** (Cheng & Sybesma, 1998; Gil, 2013). Languages with this feature are sometime called **OBLIGATORY CLASSIFIER LANGUAGES**, often referred to as **CLASSIFIER LANGUAGES** in short. Classifier languages are spoken by a large proportion of the world's population, and include Mandarin and other Chinese languages, Japanese, Korean, Austroasiatic languages, Mayan languages, and others (Aikhenvald, 2000; Gil, 2013). Below are examples with numeral classifiers in Vietnamese (3), Minagkabau (4), and Japanese (5).

(3) hai con chó
two CL dog
("two dogs") (Vietnamese, (Gil, 2013))

(4) duo ikue anjiang
two CL dog
("two dogs") (Minagkabau, (Gil, 2013))

(5) kuruma san-dai
car three-CL
("three cars") (Japanese, (Sudo, 2018))

Sortal numeral classifiers vs. Mensural numeral classifiers Researchers generally agree that there is a distinction between the two types of classifiers: Sortal numeral classifiers ((1)-(2)) categorize classes of noun entities, generally in reference to their salient perceptual properties, which are often permanently associated with the entities named by the nouns; Mensural numeral classifiers ((6)-(7)) create units to measure the quantity of the entities and usually form a more temporary relationship with the nouns (Allan, 1977; J. Tai & Wang, 1990; J. H. Tai, 1992, 1994; Cheng & Sybesma, 1998; Zhang, 2007; P. Li, Huang, & Hsiao, 2010; Gil, 2013).

(6) 两 箱 苹果
liang xiang pingguo

two box apple (“two boxes of apples”)

(7) 两 杯 水
liang bei shui
two cup water (“two cups of water”)

In this study, I only focus on sortal numeral classifiers, which are referred to here as simply as "classifiers", in context where the meaning is clear.

2.1.2 Numeral classifiers vs. grammatical gender

Classifiers are often being compared with noun classes such as grammatical gender. What is the relation between the two? Numeral classifiers are different from noun classes such as grammatical gender in size, realization, scope, and semantics (Dixon, 1986). Below is a summary of Dixon’s (1986) criteria opposing gender and classifiers.

Size Noun classes usually involve a small number of classes (2 to around 20), whereas classifier systems generally have a fair number of classifiers, with over 100 being common.

Realization Noun classes always form a closed grammatical system. For example, information about noun class may be combined in a single morpheme with definiteness (e.g., French), number (e.g., Bantu languages), or case (e.g., Latin). Classifiers, however, are always a free form.

Scope Marking of noun classes is never entirely within the noun word, instead, it also applies to other words in the sentences that may include numerals, adjectives, and sometimes is code onto the verb (e.g., Swahili). There is little variation between speakers in the use of noun classes. For classifiers, there is never any reference to a classifier outside the noun phrase. Classifier use often indicates style or mode differences, and there is variation between speakers.

Semantics Both noun classes and numeral classifiers provide means of categorization objects in the world. There is one important difference between the two classifications. A noun class affix has a fairly fixed meaning mapped to limited classes. A classifier, on the other hand, is a lexeme with more use cases and the context of use is important.

Now the distinctions between classifiers and grammatical gender are taken care of, at least for the purpose of this study, we will zoom in to classifiers in Mandarin Chinese.

2.2 A sneak peek into the history of Chinese classifiers

Here I present some properties of the Chinese classifier system arose during the historical development to help us better understand the system. Classifiers used in modern Chinese are the results of over three thousand years of language evolution. Some classifiers first appeared in the language as early as in Zhou Dynasty (around BC 1046 - BC 256), such as 匹 *pi* (CL.horse), 两（辆） *liang* (CL.vehicle), 本 *ben* (CL.book) (Huang, 1964; Da, 2004; Ma, 2015). The members of classifiers in Chinese and their usage have been evolving over time. Below I highlight the evolution of Chinese classifiers in two aspects: structural changes in syntax, and the development of members of classifier categories in semantics.

2.2.1 Structural changes

I illustrate the structural changes using two pairs of examples taken from snippets in Classical Chinese texts (originally picked by and cited in (Ma, 2015)). In each pair, the first example is older than the second example, and both examples express the same or nearly the same meaning.

The first change was from no classifiers to having classifiers. In (8), there was no classifier used for 马 *ma* (horse), and the meaning of one horse was expressed as 一

马 *yi ma* (one horse). In (9), the classifier 匹 *pi* (CL.horse) was used. The meaning meaning was expressed as 一匹马 *yi pi ma* (one CL.horse horse).

(8) 勇 请 进 观 形势， 于是 率 百 骑，各 拢 一 马 ……
Yong qing jin guan xingshi, yushi shuai bai ji, ge long **yi ma** ...
Yong request forward watch situation, then lead hundred cavalryman, each get **one horse**
"Yong requested to be allowed to move forward and watch the situation. Then he led a hundred cavalrymen. Each got on horse." （source:《水经注》卷二十八，*Shuijingzhu, Vol 28*）

(9) 勇 进 观 形势， 于是 率 百 骑，各 笼 **一 匹 马** ……
Yong jin guan xingshi, yushi shuai bai ji, ge long **yi pi ma** ...
Yong forward watch situation, then lead hundred calaryman, each get **one CL.horse horse**
"Yong moved forward and watch the situation. Then he led a hundred cavalrymen. Each got on horse." （source:《太平御览》卷八百九十七，*Taiping Yulan, Vol 897*）

The second change was from post-nominal classifier structure to pre-nominal classifier structure: Noun Numeral Classifier → Numeral Classifier Noun. In (10), several dozens of horses was expressed as 马数十匹 *ma shushi pi* (horse several-dozen CL.horse), with the noun coming first followed by the numeral and the classifier. In (11), the same meaning is expressed as 数十匹马 *shushi pi ma* (several-dozen CL.horse horse), following the word order that we normally see in modern Chinese today: Numeral Classifier Noun.

(10) 三国 时，陆逊 攻 襄阳， 于 此穴 又 得 **马 数十 匹** ……
Sanguo shi, Luxun gong Xiangyang, yu ci xue you de **ma shushi pi** ...
Sanguo time Luxun attack Xiangyang in this cave and get **horse serveral-dozen CL.horse**
"During the Thee Kingdom period, Luxun attacked Xiangyang (place name), and he again obtained several dozen horses from this cave..." （source:《北史》卷三十，*Beishi, Vol 30*）

(11) 陆逊 攻 襄阳, 又 值 此穴 中 **数十 匹** **马** 出……
Luxun gong Xiangyang, you zhi ci xue you zhong **shushi pi ma** chu ...
Luxun attack Xiangyang again encouter this cave **serveral-dozen CL.horse horse** out
"Luxun attacked Xiangyang (place name), and again he encountered several dozen horses (running) out from this cave..." （source:《册府元龟》卷三百六十四，*Cefu Yuangui, Vol 364*）

Note that the existence of these different classifier structures overlaps in time but the trend is no classifiers prior to having classifiers, and post-nominal classifiers prior to pre-nominal classifiers. Although the Numeral Noun structure is rare in modern Chinese, but this structure has not completely died out. It is preserved in some Chinese idioms such as 一马当先 *yi ma dang xian* (one horse is first), which means to take the lead. And the Noun Numeral Classifier structure is often used in the context such as cooking recipes. For example, 鸡蛋 两 个 *jidan liang ge* (egg two CL.general).

2.2.2 Semantic development of a classifier: using *tiao* as a case study

The semantics of classifiers and the categories of objects associated with classifiers evolve over time. Below I illustrate the semantic development of a classifier using 条 *tiao* (CL.strip) as a case study. Based on *Shuowen Jiezi*, an early-2nd-centry Chinese dictionary from the Han Dynasty, *tiao* originally means twigs from trees (12).

(12) 条，小枝 也。
tiao, xiaozhi ye.
"Tiao refers to twigs." (source:《说文解字》*Shuowen Jiezi*)

It is generally considered that the classifier 条 *tiao* was evolved from the noun (W. Li, 1980; Liu, 1965; Meng, 2009; Ma, 2015), extending the original meaning to two separated directions: long things (长条, *changtiao*), and itemized or organized things

(条理, *tiaoli*). The following five examples ((13) - (17)) were first cited in (Ma, 2015).

Direction 1: Typical long or slender things → atyptical long or slender things

(13) ... 但 有 小 **蛇** 数 **条**, 尽 白, 皆 杀 之 而 归。
... dan you xiao **she** shu **tiao**, jin bai, jie sha zhi er gui
.. but exist small **snake** several **CL.strip**, all white, all kill them and return

"... but there are several small snakes. They were all white. (He) killed all of them and returned." (source:《太平广记》卷四五八, *Taiping Guangji, Vol 458*)

(14) 休 想 我 系一**条** 麻布 孝 **腰裙**!
xiu xiang wo ji yi **tiao** mabu xiao **yaoqun**!
don't think I wear one **CL.strip** linen mourning **skirt**!

"Don't even think about letting me wear a linen morning dress! " (source:《荆楚臣重对玉梳记》第一折, *Jinchucheng Chongdui Yushu Ji, Act 1*)

(15) 你这 **条** **狗**, 要 多少 钱?
ni zhe **tiao gou**, yao duoshao qian?
you this **CL.strip dog**, require how-much money

"This dog of yours, how much is it? " (source:《济公全传》第八十八回, *Jigong Quanzhuan, Chapter 88*)

Direction 2: Itemized things such as laws, regulations → certain abstract things

(16) 又 增 **法** 五十**条**, 犯者徙 之西海。
you zeng **fa** wushi **tiao**, fanzhe xi zhi Xihai.
you add **law** fifty **CL.strip**, offender move to Xihai

"Fifty laws were added, and offenders were exiled to Xihai. " (source:《汉书王莽传》, *Hanshu Wangmang Zhuan*)

(17) 我 有 一**条** **计策**, 不知中 你们 意 否?
wo you yi **tiao jice**, buzhi zhong nimen yi fou?
I have a **CL.strip strategy**, don't-know fit you desire whether

"I have a strategy. I wonder whether you will like it. " (source:《水浒传》

第十六回, *Shui Hu Zhuan, Chapter 16*)

Using the classifier *tiao* as an example, we can see that the semantics of the classifier has generalized from its original meaning and the nouns that are compatible with the classifier has expanded greatly. As a result, it is quite common that the meaning and use of some parts of the classifier system become less transparent. Another way to look at the diverse use of classifier is from the perspectives of productivity and idiosyncrasy in language. Recent research on linguistic categories using Chinese classifiers suggests an exemplar-based approach provides the best account for how Chinese classifiers were extended to emerging nouns (Habibi, Kemp, & Xu, 2020).

Below I will switch gear toward toward describing and discussing productivity and idiosyncrasy in classifier use in relation to other linguistic phenomena.

2.3 Productivity, idiosyncrasy, and classifier choice

Language gives us a remarkable ability to generate novel expressions. However, unlike formal languages such as regular expressions and programming language, irregularity is pervasive in natural language. Productivity in natural language is often if not always limited to some extent (Goldberg, 1995; Yang, 2016). One of the major challenges in understanding language is to account for its partial productivity (Albright & Hayes, 2003; J. L. Bybee, 1985; Baayen, 1994, 2002; Yang, 2005, 2016).

A well-studied example is English past tense. The majority of English verbs (i.e., regular verbs) mark their past tense and past participle with the default suffix *-ed*, whereas a small number of English verbs (i.e., irregular verbs) form their past tense and past participles in other ways. There is a large body of work trying to account for the partial productivity of this system (Rumelhart & McClelland, 1986; Plunkett & Marchman, 1991; Pinker & Prince, 1988; Pinker, 1999/2011; Marcus et al., 1992; Albright & Hayes, 2003, to name a few). More recent research suggests that the problem of productivity can be seen as one of case-by-case inference in a probabilistic framework (O'Donnell, 2015). Using evidence available from English past tense (e.g., *walk* - *walked*, *go* - *went*), one can model how an agent would infer whether it is better to account for the structure using productive computation, stored items, or a combination of the two for each data point (O'Donnell, 2015).

However, to account for (partial) productivity in language, previous studies have extensively focused on morphology and more recently, multi-word expressions with binary order preferences (e.g., *bread and butter* over *butter and bread*) (Morgan & Levy, 2016). We know little about the semantic productivity of lexical categories. Exploring classifier use offers an opportunity to investigate how productivity interacts with semantic categories as each classifier is associated with at least one category of things and classifier-noun agreement is a snapshot of conceptual representation in language.

2.3.1 Productive concepts picked up by classifiers

In modern Chinese, there are about 100 active sortal classifiers (Ma, 2015). The most commonly used classifier in Mandarin Chinese is 个 *ge* (CL.general). It can be used with many nouns, including animate and inanimate ones, concrete and abstract ones. The mapping between nouns and classifiers, in most if not all cases, is not one-to-one. On one hand, a classifier can be used with more than one category of nouns, for example, 条 *tiao* (CL.strip) can be used with 鱼 *yu* (fish), 皮带*pidai* (belt), 绳子 *shengzi* (rope), 裙子 *qunzi* (dress), as well as more abstract nouns such as 通知 *tongzhi* (notification) and 信息 *xinxi* (message). On the other hand, a noun can be compatible with more than one classifier. For example, 电脑 *diannao* (computer) can be used with 台 *tai* (CL.machinery), 部 *bu* (CL.machine) or 个 *ge* (CL.general). Some nouns have a relatively strong preference for a particular classifier, whereas others split their preferences for two or more classifiers more evenly. For instance, 猫 *mao* (cat) almost always goes with the classifier 只 *zhi* (CL.animal), whereas 狗 *gou* (dog) can be used with either 只 *zhi* (CL.animal) or 条 *tiao* (CL.strip) with relatively even preference.

Below is a sample of classifiers and the categories of things they apply to based on my own native speaker knowledge and on previous work (He, 2003; Ma, 2015).

- 个 ge (CL.general): 1) General classifier for individual person or object.
- 只 zhi (CL.animal): 1) Non-human animals; 2) one of a pair, e.g., sock; 3) other small objects such as cups.
- 张 zhang (CL.flat): 1) Flat things, e.g., card; 2) furniture with a functional flat surface, e.g., table; 3) faces
- 朵 duo (CL.flower-like): 1) Flowers; 2) Flower-shape things, e.g., cloud.
- 辆 liang (CL.vehicle): 1) Vehicles.

Based on my native speaker intuition from linguistic examples, I think classifier preferences primarily live on referents, and are manifested through the use of nouns associated with the referent. Here I provide two types of examples to illustrate the nature of Mandarin classifier system as a productive set of concepts: (near) synonyms and novel nouns introduced to the language.

See the following examples (18) and (19). Both 电脑 *diannao* and 计算机 *jisuanji* refer to computers. The noun *diannao* is widely used in spoken Mandarin, and the noun *jisuanji* sounds more formal. Both can be used with the classifier *tai* and *tai* is the preferred specific classifier for both nouns.

(18) 三 台 电脑
san tai diannao
three CL.machinery computer ("three computers")

(19) 三 台 计算机
san tai jisuanji
three CL.machinery computer ("three computers")

When a novel noun is introduced, native speakers converge on the preferred classifier(s) for the noun, based on the semantics of the referent associated wit the noun. For example, when a novel word *Biyadi* was introduced, speakers know they can use the classifier *liang* (CL.vehicle) with it (20). For Mandarin-speaking listeners who did not know about *Biyadi*, they would probably be able to guess that this noun is likely to refer to a car or a truck.

(20) 李婕 买了 一辆 比亚迪。
Lijie mai-le yi liang Biyadi.
Lijie buy-ASP one CL.vehicle BYD ("Lijie bought a BYD (a car brand) car.")

Conceptually, a speaker would need to have representations of relevant properties of objects or concepts in the world in order to use classifiers appropriately. Those properties include but are not limited to animacy, shape, size, rigidness, etc. More specifically, native speakers have:

- the knowledge of how to categorize objects based on their physical properties

(21) 一条 尺子
yi tiao chizi
one CL.strip ruler ("a ruler")

(22) 一张 票
yi zhang piao
one CL.flat ticket ("a ticket")

- extensions of the knowledge from prototypical objects to other objects or abstract concepts

(23) 一条 规定
yi tiao guiding
one CL.strip regulation ("a regulation")

(24) 一条 记录
yi tiao jilu
one CL.strip record ("a record")

(25) *一张 规定
yi zhang guiding
one CL.flat regulation ("a regulation")

- knowledge of how to highlight different parts of an object using different classifiers

(26) 一朵 花
yi duo hua
one CL.flower.like flower （“one flower，probably just the flower, maybe with a stem, very unlikely with roots"）

(27) 一支 花
yi zhi hua
one CL.long.rigid flower (“one flower, with a stem")

(28) 一株 花
yi zhu hua
one CL.plant flower (“one flower with roots (the whole plant)")

- knowledge of how to disambiguate different senses of a noun using different classifiers

(29) 这 张 报纸
zhe zhang baozhi
this CL.flat newspaper （“this piece of newspaper (the paper itself)"）

(30) 这 家 报纸
zhe jia baozhi
this CL.enterprise newspaper (“this newspaper (as a press)")

Idiosyncratic preferences in classifier use Some classifiers are more transparent than others whereas some are semantically opaque and the organization of categories is complex (J. H. Tai, 1994; Zhang, 2007; Gao & Malt, 2009; Srinivasan, 2010; Ma, 2015). However, native speakers of a classifier language have a clear and shared intuition about which classifier(s) to use for a new object (Allan, 1977; Hiranburana, 1979; P. Li et al., 2010). Classifier preferences are largely predictable from the concepts picked out by the classifiers, but not entirely.

For example, *shu* (book) has a strong preference toward the classifier *ben* (CL.book-like) over the general classifier *ge* (CL.general). The use of *ge* (CL.general) with *shu* (book) is generally considered quite odd (see (31) and (32)).

(31) 一本 书
yi ben shu
one CL.book-like book ("one book")

(32) *一个 书
yi ge shu
one CL.general book ("one book")

However, both *ben* (CL.book-like) (31) and *ge* (CL.general) (34) sound fine for other book-like objects such as *jishiben* (notepad).

(33) 一本 记事本
yi ben jishiben
one CL.book-like notepad ("one notepad")

(34) 一个 记事本
yi ge jishiben
one CL.general notepad ("one notepad")

The oddness of (32) is likely partially rising from STATISTICAL PREEMPTION, where indirect negative evidence comes from hearing one form A, in a context where another form B is expected (Goldberg, 2016). In this case (31), *ge* (CL.general) is preempted by *ben* (CL.general). These features in language are ubiquitous and have been well studied in other areas of language, such as morphology (Aronoff, 1976; Kiparsky, 1993) and argument structure (E. V. Clark, 1987; Goldberg, 1995, 2006, 2011, 2016).

Another source of idiosyncrasy comes from semantically opaque categories exemplified by some members of a classifier. For example, the classifier *tiao* (CL.strip) is compatible with dog (35), but not with cats (36).

(35) 两 条 狗
liang tiao gou
twp CL.strip dog ("two dogs")

(36) *两 条 猫
liang tiao mao
twp CL.strip cat ("intended: two cats")

However, unlike the example of *ben* in (31), the classifier *tiao* does not preempt *gou* (dog) completely. There is another specific classifier 只 *zhi*(CL.animal) that is compatible with both dogs (37) and cats (38).

(37) 两 只 狗
liang zhi gou
twp CL.animal dog （“two dogs")

(38) 两 只 猫
liang zhi mao
twp CL.animal cat （“two cats")

As I mentioned in 2.2.2, the Chinese classifier system has gone through thousands of years of evolutions and it keeps evolving. Some of the idiosyncratic traits we found in the classifier system today might be a combination of historical, cultural, and cognitive influences that are still less clear to us. Chinese classifiers provide a distinctive opportunity to study the nature of productivity and idiosyncrasy of language in a domain with rich conceptual structures. While I certainly cannot solve all the questions in the entire classifier system in Mandarin Chinese within this thesis, I acknowledge the complexity of the system, and focus on the cognitive aspect of this problem with a goal of investigating theories of speaker choice from the perspective of language production.

2.3.2 Speaker choice investigated in this study

While different nouns are compatible with different SPECIFIC classifiers, the GENERAL classifier *ge* (CL.general) can be used with most nouns, as illustrated in Table 2.1. Often, the use of a general versus a specific classifier for a given noun carries little or no meaningful distinction for the utterance, as shown in (39) and (40).

(39) 我 买了 一台 打印机
wo mai-le yi tai dayinji
I buy-ASP one CL.machinery printer (“I bought a printer")

(40) 我 买了 一个 打印机
wo mai-le yi ge dayinji
I buy-ASP one CL.general printer (“I bought a printer")

In this study, I will focus on speaker choice in classifier production to the binary choice between using the general classifier *ge* (CL.general) and a specific classifier (CL.specific) in contexts where the two instances express more or less the same meaning.

Considering the connection between productivity/idiosyncrasy and word frequency, it is natural to question whether this could create a potential confound for the studies: if classifier-noun preferences can be idiosyncratic, the we would expect frequent nouns would be more likely to hold these idiosyncratic preferences for specific classifiers. This is a valid concern and I will offer three treatments and a perspective to address this issue.

First, using a specific classifier does not equal exhibiting idiosyncrasy. Most if not all specific classifiers have at least one category that is semantically productive in some way. For example, the classifier *liang*(CL.vehicle) is semantically productive and applies to all kinds of vehicle. Second, in our empirical work, we did not include frequent nouns such as *shu* (book) that are only compatible with a specific classifier (e.g., *ben* (CL.book) in (31)) and not with the general classifier (32)[1]. By excluding nouns with such statistical preemption with regard to classifier use, we can avoid a large part of the potential cofound. Third, in the corpus work, we controlled for noun frequency and to preview the results, it is noun surprisal, not noun frequency that crucially predicts classifier choice. In our picture naming experiments, we found effects of time pressure and number of objects, and these effects are not predicted by the productivity/idiosyncrasy account. Finally, instead of viewing word frequency as a potential confound, another way to look at it is to think of word frequency as an indicator of linguistic experience of the language users. The more linguistic experience the speaker has with a noun, the more available it might become when it

[1]Based on the author’s native speaker intuition and consultation of the Beijing Language and Culture University Corpus (BLCU Center, BCC) Corpus (Xun, Rao, & Zang, 2016).

	个 ge CL.general	张 zhang CL.flat	幅 fu CL.frame	项 xiang CL.item	名 ming CL.person	...
teacher	✓				✓	
table	✓	✓				
painting	✓	✓	✓			
project	✓			✓		
...						

Table 2.1: Matrix of Classifier-Noun compatibility. A checkmark indicates the classifier is compatible with the noun.
dayinji

comes to speaking. And this gives us an measure for availability in lexical access in picture-naming without sentential context.

2.4 Predictions of language production theories for Chinese classifiers

Below I illustrate the predictions of Uniform Information Density and Availability-based Production on the relationship between noun predictability and classifier choice using the above examples ((39) - (40)).

A specific classifier and a general classifier have different effects on the information densities of the following nouns. A specific classifier is more likely to reduce the information density of the upcoming noun than a general classifier because a specific classifier constrains the space of possible upcoming nouns more tightly (Klein, Carlson, Li, Jaeger, & Tanenhaus, 2012). In (39), when comprehenders get to the specific classifier *tai* (CL.machinery), they would know from the specific classifier that the upcoming noun is likely to be something related to machines. When they get to the noun *dayinji* (printer), the surprisal of the noun would be reduced due to the presence of the specific classifier (as opposed to the general classifier). In (39) in contrast, the general classifier *ge* (CL.general) provides little information about the upcoming noun: it could be a lot of things. When they get to the noun *dayinji* (printer), it would contain high surprisal. If the goal is to convey information at a relatively constant rate, one strategy is to use a specific classifier in cases when the noun would otherwise have high surprisal. Thus the UID hypothesis predicts that speakers will be more likely to choose a **specific** classifier when the noun predictability would otherwise be low. In other words, the less predictable the noun, the more likely speakers will choose a specific classifier. This is because the use of a specific classifier offloads some of the information from the noun, making the information content less dense at the noun, potentially avoiding a peak in information density and facilitating comprehension.

Availability-based Production, on the other hand, makes different predictions than the UID account, provided two plausible assumptions. First, I assume the speaker must access a noun lemma in order to access its appropriate specific classifier. Second, I assume less predictable nouns are harder and/or slower to access. Under these assumptions, when the noun *dayinji* (printer) is accessed quickly, the speaker is more likely to access and produce an appropriate specific classifier (e.g., *tai* (CL.machinery)) associated with the noun. If the noun *printer* (printer) and its associated specific classifier candidates are accessed slowly, the speaker is likely to use the general classifier *ge* (CL.general) as a default option. The ABP hypothesis thus predicts that the less predictable the following noun, the more likely speakers will be to choose a **general** classifier. Note that this is not the only possible version of an availability-based theory as applied to Chinese classifiers[2], but it is a crucial version here, because it makes opposite predictions to UID.

Here I consider the general classifier *ge* to be available quickly for three reasons. First, the general classifier *ge* is compatible with most nouns and it is the most productive classifier in Mandarin Chinese. Second, *ge* is ranked the 8th most frequently used word in modern Chinese (Xiandai, 2008). Third, the character *ge* (identical as the word *ge*) is frequently used in a wide variety of contexts, such as 这个 *zhege* (“this"), 那个 *nage* (“that"), 个体 *geti* (“individual").

To preview the results, I found that the distribution of speaker choice in classifier

[2]Another logically possible version could be this: if the speaker need to access the classifier in order to produce a noun, speaks would use specific classifiers to facilitate noun production.

production supports the speaker-oriented Availability-based account.

Chapter 3

Comparing Theories of Speaker Choice in Naturalistic Language Datasets

3.1 Introduction

The expressivity of natural language often gives speakers multiple ways to convey the same meaning. Meanwhile, linguistic communication takes place in the face of environmental and cognitive constraints. For instance, language users have limited memory and cognitive resources, the environment is noisy, and so forth. What general principles govern speaker choice in the face of alternations that are (nearly) semantically invariant? To the extent that we are able to provide a general answer to this question it will advance our fundamental knowledge of human language production.

Studies have shown that speaker choices are very often sensitive to contextual predictability. For well-studied cases of optional reduction in language, the following rend is widespread: the more predictable a linguistic unit is, the more likely it is to get reduced. Predictable words are phonetically reduced (Jurafsky et al., 2001; Bell et al., 2009; Seyfarth, 2014) and tend to have shorter lexical forms (Piantadosi et al., 2011), and optional function words are more likely to be omitted when the phrase they introduce is predictable (Jaeger, 2010). Yet it is unclear to what extent speakers' choices when faced with an alternation are made due to audience design or to facilitate production.

For example, the above pattern of predictability effects in optional reduction phenomena is predicted by both the Uniform Information Density (UID) hypothesis (Levy & Jaeger, 2007; Jaeger, 2010), a theory which that the speaker aims to convey information at a relatively constant rate and which can be motivated via considerations of optimality from the comprehender's perspective (Smith & Levy, 2013), and by the speaker-centric availability-based production hypothesis (K. Bock, 1987; V. S. Ferreira & Dell, 2000), which hypothesizes that the dominant factor in determining speaker choice is that the speakers uses whatever material is readily available when it comes time to convey a particular part of a planned message.

Here we argue that for a different optionality phenomenon, namely classifier choice in Mandarin Chinese, UID and ABP make opposite predictions regarding the relationship between the predictability of upcoming material and speaker choice. In a corpus analysis of Mandarin Chinese, we show that the distribution of speaker choices supports the availability-based production account.

3.2 Uniform Information Density and Availability-based Production

In Sections 2 and 3, we explain why the UID and availability-based production accounts make the same predictions in many cases, but can be potentially disentangled using Chinese classifier choice. Here we exemplify predictions of these two accounts in the case of optional function word omission.

For optional function word omission such as *that*-omission ((1) and (2)), predictability effects have been argued to be consistent with both the speaker-oriented account of AVAILABILITY-BASED PRODUCTION (K. Bock, 1987; V. S. Ferreira & Dell, 2000) and the potentially audience-oriented account of UNIFORM INFORMA-

TION DENSITY (Levy & Jaeger, 2007). On both accounts, but for different reasons, the less predictable the clause introduced by the functional word, the more likely the speaker will be to produce the function word *that*.

(1) The student (that) you tutored graduated.

(2) The woman thought (that) we were crazy.

The UID hypothesis claims that within boundaries defined by grammar, when multiple options are available to encode a message, speakers prefer the variant that distributes information density most uniformly, thus lowering the chance of information loss or miscommunication (Levy & Jaeger, 2007; Jaeger, 2010). In (1), if the function word *that* is omitted, the first word of the relative clause *you* serves two purposes: signaling the onset of the relative clause, and conveying part of the contents of the relative clause itself. These both contribute to the information content of the first relative clause-internal word. If one or both is high-surprisal, then the first relative clause-internal word might be a peak in information density, as illustrated in Figure 3-1 (top left). If instead the function word *that* is produced, *that* signals the onset of the relative clause, and *you* only communicates part of the content of the relative clause itself. This could help eliminate any sharp peak in information density, as illustrated in Figure 3-1 (bottom left). Thus, if the speaker's goal is to transfer information as smoothly as possible, the less predictable the upcoming clause, the more inclined the speaker would be to produce the function word *that*.

On the availability-based production hypothesis, speaker choice is governed by the relationship by the relative time-courses of (i) when a part of a message needs to be expressed within an utterance, and (ii) when the linguistic material to encode that part of the message becomes available for production. If material that specifically encodes a part of the message is available when it comes time to convey that part of the message, it will be used—that is the PRINCIPLE OF IMMEDIATE MENTION of (V. S. Ferreira & Dell, 2000). If, on the other hand, that material is not yet available, then other available material consistent with the grammatical context produced thus far and that does not cut off the speaker's future path to conveying the desired content will be used. In (1), assuming the function word *that* is always available when the speaker plans to produce a relative clause, the speaker will produce *that* when the upcoming relative clause or the first part of its contents are not yet available. If phrase structures and phrase contents take longer to become available when they are lower-predictability—an assumption consistent with the literatures on picture naming (Oldfield & Wingfield, 1965) and word naming (Balota & Chumbley, 1985)—then the less predictable the relative clause, the lower the probability that its first word, w_1, will be available when the time comes to begin the relative clause, as illustrated in Figure 3-2 (left). Under these circumstances, the speaker would choose to produce other available material, namely function word *that*. If, in contrast, the upcoming relative clause is predictable, then w_1 will be more likely to be available, and the speaker would be more likely to omit the function word *that* and immediately proceed with w_1.

While these two accounts differ at many levels, they make the same prediction for function word omission in syntactic reduction such as (1) and (2). It is difficult to disentangle these accounts empirically.[1] Below we will show that for a different optionality phenomenon, namely classifier choice in Mandarin, these accounts may make different predictions.

[1] Prior work (Jaeger, 2010) acknowledged this entanglement of the predictions of these accounts, and attempted to tease the accounts apart via joint modeling using logistic regression. The present study builds on these efforts by exploring a case involving a starker disentanglement of the accounts' predictions.

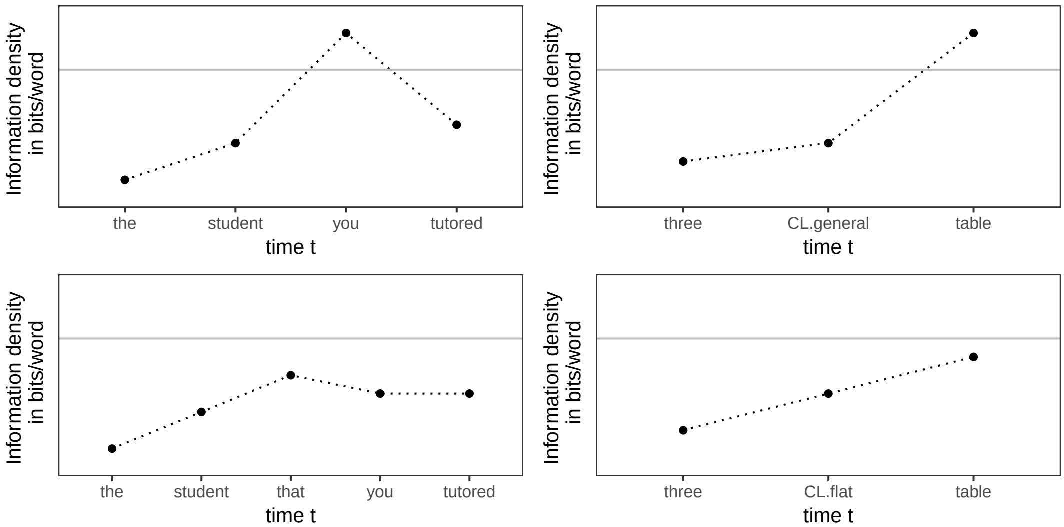

Figure 3-1: Schematic illustrations of Uniform Information Density in the context of relative clause (left) and classifier choice (right). The grey lines indicate a hypothetical channel capacity.

3.3 Classifiers in Mandarin Chinese

Languages in the world can be broadly grouped into classifier languages and non-classifier languages. In non-classifier languages, such as English and other Indo-European languages, a numeral modifies a noun directly: e.g., *three tables, two projects*. In Mandarin Chinese and other classifier languages, a numeral classifier is obligatory when a noun is to be preceded with a numeral (and often obligatory with demonstratives): e.g., *san zhang zhuozi* “three CL.flat table", *liang xiang gongcheng* “two CL.item project". Although it has been hypothesized that numeral classifiers play a functional role analogous to that of the singular–plural distinction in other languages (Greenberg, 1972), it is not clear whether there is a meaningful correlation between the presence of numeral classifiers and plurality among the languages of the world (Dryer & Haspelmath, 2013).

In Mandarin, classifiers, together with their associated numeral or demonstrative, precede the head noun of a noun phrase. There are about 100 individual numeral classifiers (Ma, 2015). While different nouns are compatible with different SPECIFIC classifiers, there is a GENERAL classifier *ge*(个) that can be used with most nouns. In some cases, the alternating options between using a general or a specific classifier

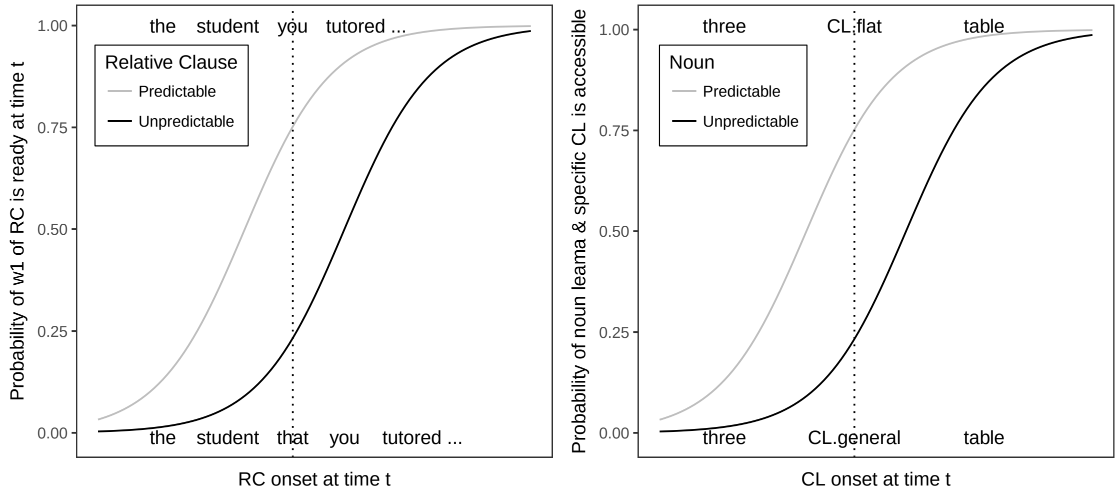

Figure 3-2: Schematic illustrations of availability-based production in the context of relative clause (left) and classifier choice (right). X axis presents the progression of time. The dotted lines indicate onset times for relative clause and classifier respectively.

with the same noun are almost semantically invariant. Table 3.1 shows examples of classifier options in fragments of naturally occuring texts.

Yet these options have different effects on the information densities of the following nouns. A specific classifier is more likely to reduce the information density of the upcoming noun than a general classifier because a specific classifier constrains the space of possible upcoming nouns more tightly (Klein et al., 2012). Consider the following pair of classifier examples (3) and (2).

(3) 我 买了 三 张 桌子
wo mai-le san zhang zhuozi
I bought three CL.flat table ("I bought three tables")

(4) 我 买了 三 个 桌子
wo mai-le san ge zhuozi
I bought three CL.general table ("I bought three tables")

As shown in Figure 3-1 (top right), while a general classifier has some information (e.g., signaling there will be a noun), it has relatively low information density—it is the most frequent and generally the highest-probability classifier in many contexts. In comparison, as illustrated in Figure 3-1 (bottom right), a specific classifier has higher information density—specific classifiers are less frequent than the general classifier and typically lower-predictability—but, crucially, it constrains the hypothesis space for the identity of the upcoming noun, since the noun's referent must meet certain semantic requirement that the classifier is associated with. The UID hypothesis predicts that speakers choose a **specific** classifier more often when the predictability of the noun would otherwise be low.

Availability-based production, provided three plausible assumptions, makes different predictions than UID. The first assumption is that a speaker must access a noun lemma in order to access its appropriate specific classifier. The second assumption is that unpredictable noun lemmas are harder and/or slower to access (as described in Section 3.2, this assumption is supported by findings from the naming literature). The third assumption is that the general classifier is *always* available, regardless of the identity of the upcoming noun, as it is the most frequently used classifier and is compatible with a lot of nouns. Under these assumptions, for unpredictable nouns, specific classifiers will less often be available to the speaker when the time comes to initiate production of classifier, as shown in Figure 3-2 (right). Since noun lemmas need to be accessed before their associated specific classifiers, the less predictable the noun, the less likely the noun lemma and hence the associated specific classifier is to be available by the classifier onset time t. The general classifier, in contrast, is always accessible. Under these assumptions, the availability-based production hypothesis thus predicts that speakers choose a **general** classifier more often when the following noun is less predictable.

3.4 Creating a corpus of classifier-noun pairs

To provide data for this study, we created a corpus of naturally occurring classifier-noun pairs from SogouCS, a collection of online news texts from various channels of

Noun	个 (ge, CL.general)	项 (xiang, CL.item)	张 (zhang, CL.flat)
公告 announ-cement	一 口 气 发布 11 个 公告 a CL breath release 11 CL announcement "release 11 announcements at one go"	连续 发布 三 项 公告 consecutively release three CL announcement "release three announcements in a row"	门口 贴了 一 张 公告 door paste a CL announcement "there is an announcement on the door"
账单 bill	女儿 拿着 一 个 账单 就 过来了 daughter carry a CL bill at once come "daughter came with a bill at once"	not co-occurring	在 一 张 账单 上 解决 所有 收费 问题 on a CL bill solve all charge problem "solve all charge problems on a bill"
工程 project	跟 圆明园 有关 的 一 个 工程 to Yuanmingyuan related de a CL project "a project related to Yuanmingyuan"	抓好 六 项 重点 工程 grasp six CL key project "manage six key projects"	not co-occurring
活动 activity	昨天 我 参加了 一 个 活动 yesterday I attend a CL activity "yesterday I attended an activity"	广州市 今天 开展 的 一 项 活动 Guangzhou today hold de a CL activity "an activity held by Guangzhou today"	not co-occurring

Table 3.1: Examples from development set of available classifier options that are semantically (near-)invariant

Sohu News (Sogou, 2008). [2]. The deduplicated version of the corpus has 11,548,866 sentences. To parse and annotate the data, we built a pipeline to 1) clean and deduplicate the data, 2) part-of-speech tag and syntactically parse the clean text, and 3) extract and filter classifier-noun pairs from the parsed text. Although an online news corpus is not ideal for investigating speaker choice, we use this corpus as a first attempt to approximate classifier use, due to its size and the ensuing statistical power of analyses using it.

3.4.1 Cleaning and deduplication

Since the data contain web pages, many snippets are not meaningful content but automatic page generation such as legal notices. To use this corpus as a reasonable approximation of language experience of speakers, we performed deduplication on the data, following similar practice adopted by other work dealing with web-based corpora (Buck, Heafield, & Van Ooyen, 2014). After cleaning the text, we removed repeated lines in the corpus.

[2]The code for processing the texts and creating the corpus of classifier-noun pairs is available at `https://github.com/meilinz/yitiao`. The processed data and analysis code are available at `https://osf.io/8zwrm/`

3.4.2 Word segmentation, POS-tagging and syntactic parsing

We used the Stanford CoreNLP toolkit for word segmentation, part-of-speech tagging, and syntactic parsing (Manning et al., 2014). We used the Shift-Reduce model in parsing (Zhu, Zhang, Chen, Zhang, & Zhu, 2013). We also obtained dependency parsing results as part of the Stanford CoreNLP output.

3.4.3 Extracting and filtering classifier-noun pairs

From the parsed corpus, we extracted all observations where the head noun has a `nummod` relation with a numeral and the numeral has a `mark:clf` relation with a classifier. 3-3 illustrates two of such examples. We included classifiers in the list of 105 sortal numeral classifiers documented in previous research (Ma, 2015) that were identified by the Stanford CoreNLP toolkit. For the purpose of gaining more or less semantically invariant alternation, we excluded classifiers such as *zhong*(CL.kind) that make a clear truth-conditional semantic contribution to the noun phrase.

(5) liang ge xiangjiao
two CL.general banana
"two bananas"

(6) liang zhong xiangjiao
two CL.kind banana
"two kinds of bananas"

We did further filtering to get nouns that can be used with both the general classifier and at least one specific classifier. This left us 1,479,579 observations of classifier-noun pairs.

To construct the development set[3], we randomly sampled about 10% of the noun types (1,179) and extracted all observations with these noun types. We manually checked and filtered applicable classifiers for these noun types and we ended up with 713 noun types for the development set. For the test set, we also randomly sampled about 10% of the noun types (1,093) and extracted all observations with these noun

[3] Note that we do not have a training set for the classifier-noun pairs.

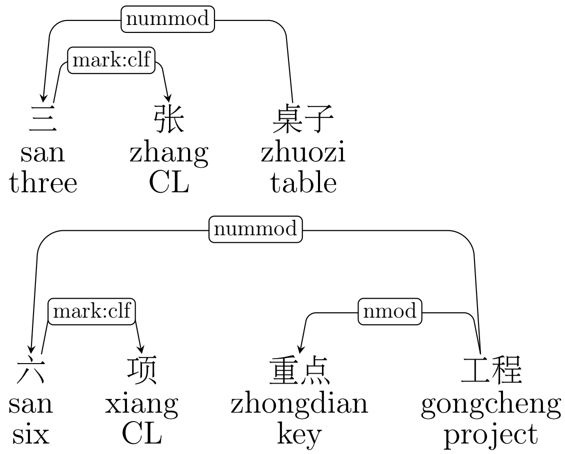

Figure 3-3: Classifier examples where the head noun has a nummod relation with a numeral and the numeral has a mark:clf relation with the classifier

types. We did not perform manual filtering of the test set. We reserve the remaining 80% of the data for future work.

3.5 Model estimation

We use SURPRISAL, the negative log probability of a word in the context (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; S. L. Frank & Bod, 2011; Smith & Levy, 2013), generated from a language model to estimate noun predictability. Because classifiers occur before their corresponding nouns in our target sequences (`Numeral Classifier Noun`), to avoid circularity, we mapped all target classifiers to an identical tag `CL` in the segmented text for language modeling, analogous to the procedure used Levy and Jaeger (2007) and similar studies. We implemented 5gram modified Kneser-Ney smoothed models with the SRI Language Modeling toolkit (Stolcke, 2002) and performed ten-fold cross-validation with the SogouCS corpus (Sogou, 2008) to estimate noun surprisal.

We used a mixed-effect logit model to investigate the relationship between noun predictability and classifier choice. The dependent variable was the binary outcome of whether a general or a specific classifier was used. For each noun type, we also identified its most frequently used specific classifier. We included two predictors in the analysis: noun surprisal and log noun frequency.[4] We included noun frequency as a control factor for two reasons. First, noun frequency has shown effects on many aspects of speaker behavior. Second, surprisal and frequency of a word are intrinsically correlated. Taken together, these two reasons make noun frequency an important potential confound to be controlled for in investigating any potential effect of noun surprisal on classifier choice.

We included noun and potential specific classifier as random effects, both with random intercepts and random slopes for noun surprisal. This random effect structure is maximal with regard to testing effects of noun surprisal (Barr, Levy, Scheepers, & Tily, 2013), which varies within noun and within classifier. We then applied the model to the test set. The full formular in the style of `R`'s `lme4` package is:

```
classifier_choice ~ noun-surprisal + log_noun_frequency
+ (1 + noun_surprisal | noun )
+ (1 + noun_surprisal | potential_specific_classifier)
```

We used Markov chain Monte Carlo (MCMC) methods in the `R` package `MCMCglmm` (Hadfield, 2010) for significance testing. We based our p-values on the posterior distribution of regression model parameters using an uninformative prior and determining the largest possible symmetric posterior confidence interval on one side of zero, as is common for MCMC-based mixed model fitting (Baayen, Davidson, & Bates, 2008).

3.6 Results

In both the development set and the test set, overall we saw more observations with a specific classifier than with a general classifier (55.4% vs. 44.6% in the development set, 63.1% vs. 36.9% in the test set). For the development set, we find that the less predictable the noun, the less likely a specific classifier is to be used ($\beta = -0.038$, $p < 0.001$, Figure 3-4). There was no effect of noun frequency ($\beta = 0.018$, $p = 0.51$, Figure 3-5). For the test set, the result of noun predictability replicates ($\beta = -0.059$, $p <$

[4]We used base 2 here to be consistent wit the base used in noun surprisal.

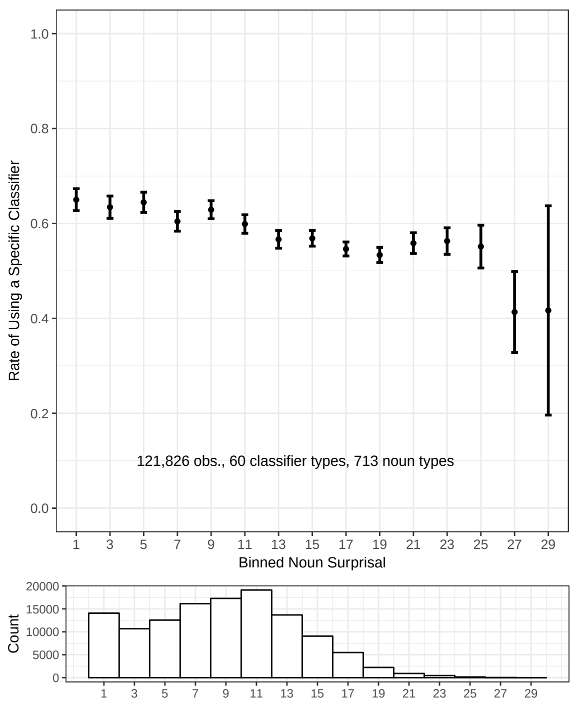

Figure 3-4: Dev set: N-gram estimated noun surprisal
and the rate of using a specific classifier (as opposed
to the general classifier *ge*).

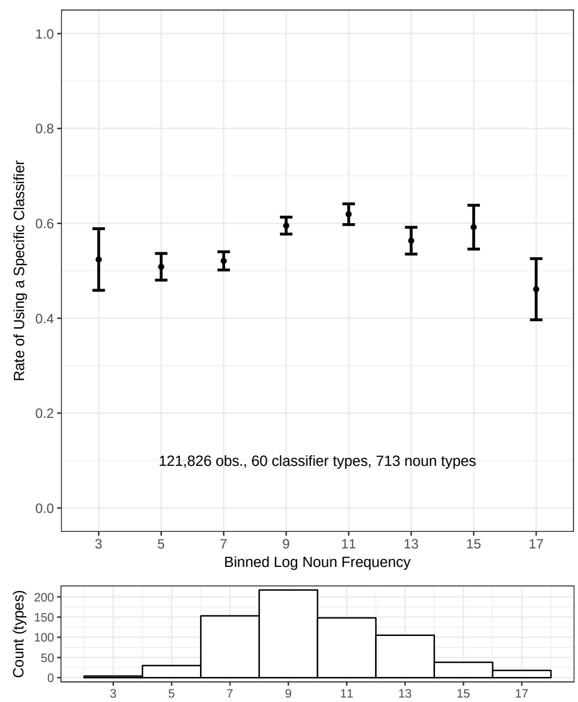

Figure 3-5: Dev set: Noun frequency (log scale) and
the rate of using a specific classifier (as opposed to
the general classifier *ge*).

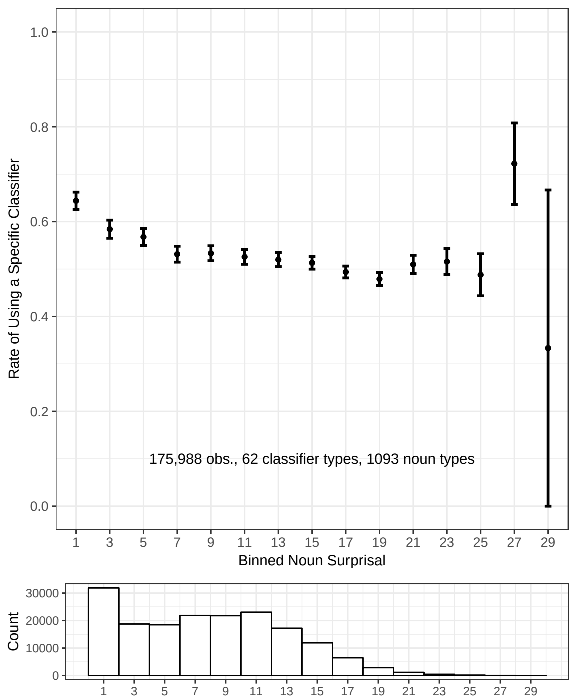

Figure 3-6: Test set: N-gram estimated noun surprisal
and the rate of using a specific classifier.

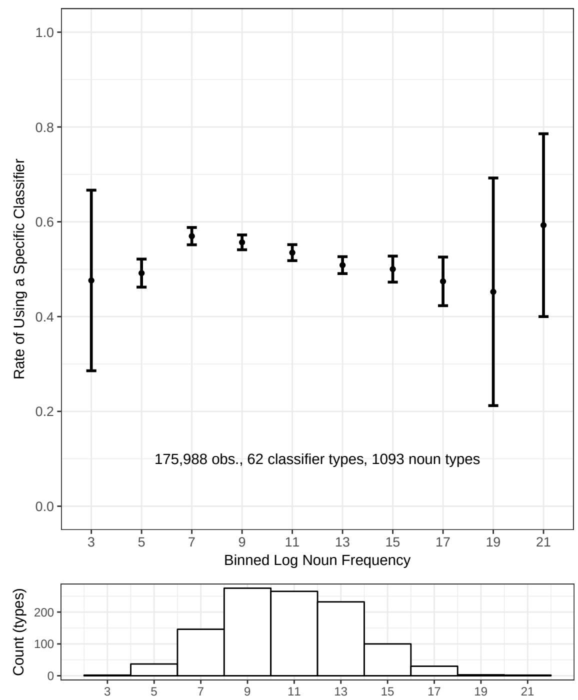

Figure 3-7: Test set: Noun frequency (log scale) and
the rate of using a specific classifier.

0.001, Figure 3-6).[5] In the test set but not in the development set, we also found an effect of noun frequency ($\beta = -0.11,\ p < 0.001$, Figure 3-7): the more frequent the noun, the less likely a specific classifier is to be used. Further analysis suggests that this effect of noun frequency in the test set is likely to be an artifact of incorrect noun–classifier associations that would disappear were we to filter the test set in the same way as we filtered the development set.[6] The consistent effect of noun surprisal on classifier choice in both our development and test sets supports the availability-based production hypothesis, and is inconsistent with the predictions of UID.

[5] As can be seen in Figure 3-6, there is a bump at bin 27 in the rate of using a specific classifier. We consider this likely to be due to data sparsity: the number of observations is small in the last two bins of noun surprisal ($n = 27$ and $n = 3$), and there is no such bump in the development set.

[6] We found a marginal effect of noun frequency in our unfiltered development set, where the more frequent the noun was, the less likely it was used with a specific classifier. We did further analysis with the dev set and found that the“nouns" (some of them were misclassified as nouns from the results of the automatic parsing) that were excluded tend to have a higher frequency compared to the ones that were included, and the excluded ones also had a lower rate of concurring with a specific classifier. This tendency suggests that in the unfiltered test set, illegible nouns may contribute at least partially to the noun frequency effect.

One potential concern regarding the above conclusion that noun predictability drives classifier choice is that it might not fully take into account effects of the frequencies of classifiers themselves on availability. The availability-based production hypothesis does not exclude the possibility that a classifier's accessibility is substantially dependent on its frequency, and the general classifier is indeed the most frequently used classifier. However, if specific classifier frequency were confounding the apparent effect of noun surprisal that we see in our analysis, there would have to be a correlation in our dataset between specific classifier frequency and noun surprisal. Our inclusion of a by-specific-classifier random intercept largely rules out the possibility that even a correlation that the above-mentioned one could be driving our effect. To be thorough, we tried a version of our regression analysis that also include a fixed effect for the log frequency of potential specific classifier as a control. We did not find any qualitative change to the results: the effect of noun surprisal on specific classifier choice remains the same. We also note that in this new analysis, we do not find a significant effect of specific classifier log frequency on classifier choice ($p = 0.629$ for the dev set and $p = 0.7$ for the test set). This additional analysis suggests that it is unlikely that the effect of specific classifier frequency to be driving the effect of noun surprisal.

Overall, we did not find evidence for the UID hypothesis at the level of alternating options with different information density, in our case, a specific classifier versus a general classifier. We demonstrate that within the scope of near semantically invariant alternation, classifier choice is modulated by noun predictability with the tendency to facilitate speaker production. Our results lend support to an availability-based production model. We did not find consistent evidence for the effect of noun frequency on classifier choice. The effect of noun frequency remains unclear and we will need to test it with a larger sample of noun types.

3.7 Conclusion

Though it has proven difficult to disentangle UID and availability-based production through optional word omission phenomena, we have demonstrated here that the two accounts can potentially be distinguished through at least one word alternation phenomenon. The UID hypothesis predicts that predictable nouns favor the *general* classifier whereas availability-based production predicts that predictable nouns favor a *specific* classifier. Our empirical results favor the availability-based production account.

To the best of our knowledge, this is the first study that demonstrates contextual predictability is correlated with classifier choice. This study provides a starting point to understand the cognitive mechanisms governing speaker choices as manifested in various language optionalities. Ultimately we plan to complement our corpus analysis with real-time language production experiments to more throughly test hypotheses about speaker choice.

Chapter 4

Availability-Based Production Predicts Speakers' Real-time Choices of Mandarin Classifiers

4.1 Introduction

The simple act of speaking may typically seem effortless, but it is extraordinarily complex. Speakers must plan the message they wish to convey, choose words and constructions that accurately encode that message, organize those words and constructions into linearly-sequenced utterances, keep track of what has been said, and execute each part of their speaking plans at the correct time. Throughout this process, speakers face choices in structuring their intended message into an utterance. One central question for a computationally precise theory of language production is thus: When multiple options are available to express more or less the same meaning, what general principles govern a speaker's choice? To what extent do speakers make choices that potentially facilitate comprehenders, and to what extent do they make choices that are preferable from a production standpoint? Here we approach these questions from the standpoint of contextual predictability, which is known to affect a wide range of speaker choices. Specifically, we investigate the influence of contextual predictability on the encoding of linguistic content manifested by speaker choice in a classifier language. Two major theories of sentence production, Availability-Based Production (ABP; K. Bock, 1987; V. S. Ferreira & Dell, 2000) and Uniform Information Density (UID; Levy & Jaeger, 2007; Jaeger, 2010), make conflicting predictions about the distribution of speaker choices when more than one classifier could be used in a given context. We report a language production experiment on classifier choice that adjudicates between these theories.

In languages with a grammaticalized count–mass distinction, such as English, count nouns such as *table* can be used with a numeral directly and typically exhibit a singular–plural morphological marking (e.g., *one table, three tables*), whereas mass nouns such as *sand* cannot co-occur with numerals directly without some kind of measure word (e.g., *three cups of sand*) and do not have a plural morphology on the noun (e.g., **three sands*). In classifier languages such Mandarin, in contrast, nouns lack obligatory singular–plural morphological marking and cannot directly co-occur with numerals. Instead, a numeral classifier is required when a noun is modified by a numeral or a demonstrative. Linguists generally agree that there is a distinction between two types of Chinese classifiers: count classifiers, which we focus on here, and mass classifiers (J. H. Tai, 1994; Cheng & Sybesma, 1998; P. Li, Barner, & Huang, 2008).[1] Among count classifiers, which are used with nouns that denote individuals or groups of individuals, different SPECIFIC classifiers are compatible with different nouns, but the GENERAL classifier *ge* (个) can be used with almost any noun. Often, the choice of general versus specific classifier for a given noun carries little to no meaning distinction for the utterance, as illustrated in (1) and (2) below.

(1) 我 卖了 三 台 电脑
wo mai-le san **tai** diannao
I sold three CL.machinery computer ("I sold three computers")

[1] A count classifier (e.g., two CL.top hat ("two hats")) is used to categorize a class of noun entities in reference to their salient perceptual properties, which are often permanently associated with the entities named by the class of nouns. A mass classifier (e.g., two box (of) hat ("two boxes of hats")) creates a unit and form a temporary relationship with the noun. Because using different mass classifiers often change the semantics of the noun phrase, here we only focus on count classifiers (henceforth, classifiers).

(2) 我 卖了 三 个 电脑
wo mai-le san **ge** diannao
I sold three CL.general computer (“I sold three computers")

In this study, we focus on speaker choice between general and specific count classifiers for nouns where both options convey more or less the same meaning. When the upcoming noun is unpredictable, a specific classifier would constrain the range of possible nouns more than the general classifier, thus increasing the predictability of the upcoming noun and potentially benefiting comprehension. The Uniform Information Density account thus predicts that speakers will prefer specific classifiers for unpredictable nouns. However, Availability-Based Production predicts that the specific classifier may be dispreferred from a production standpoint if the general classifier is more easily available. Which of these two accounts better predict classifier choice in real-time production? In other words, does noun predictability affect classifier choice, and if so, in which direction? Here we use a picture-naming experiment to address this question.

Before diving into the experiment, we first briefly introduce why we focus on predictability effects and how the two accounts predict speaker choices with regard to optional reduction in language.

4.2 Predictability Effects on Optional Reduction

It has been shown that contextual predictability plays a role in optional reduction in language, where more predictable content tend to yield a greater rate of reduction in the linguistic form. At the lexical level, predictable words are phonetically reduced (Jurafsky et al., 2001; Bell et al., 2009; Seyfarth, 2014) and tend to have shorter forms (Piantadosi et al., 2011; Mahowald et al., 2013). At the syntactic level, optional function words are more likely to be omitted when the phrase they introduce is predictable (Levy & Jaeger, 2007; Jaeger, 2010). For example, in English relative clauses (henceforth RCs) such as (3), speakers can but do not have to produce the relativizer *that*. We refer to the omission of *that* as OPTIONAL REDUCTION.

(3) I created a mobile app dancers like.

(4) I created a mobile app that dancers like.

For optional function word omission, predictability effects have been argued to be consistent with both the speaker-oriented account of Availability-Based Production, where the speaker mentions material that is readily available first, and the potentially audience-oriented account of Uniform Information Density, where the speaker aims to convey information at a relatively constant rate. These two accounts have proven difficult to disentangle empirically. For different reasons, both accounts predict that the less predictable the clause introduced by the function word, the more likely the speaker would be to produce the function word *that*.

4.2.1 Uniform Information Density

Uniform Information Density proposes that within boundaries defined by grammar, when multiple options are available to express the message, speakers prefer the variant that distributes information density more uniformly throughout the utterance, to lower the chance of information loss or miscommunication (Levy & Jaeger, 2007; Jaeger, 2010). Multiple formalizations are possible under this account (Genzel & Charniak, 2002; Aylett & Turk, 2004; Maurits et al., 2010; Levy, 2018).

In (3), where the relativizer *that* is omitted, the first word of the relative clause w_1 (*dancers* in this case) is highly unpredictable and would convey two pieces of information: both the onset of the relative clause and part of the content of the relative clause itself. These both contribute to the information content of w_1, which can be measured using SURPRISAL, the negative log-probability of the word in context (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013). In (4), having *that* at the onset of the RC splits these two pieces of information apart, offloading the relative clause's onset onto *that* so that *dancers* only conveys relative clause-internal content and thus has lower information content, potentially avoiding a peak in information density and thus facilitating comprehension.

4.2.2 Availability-Based Production

Availability-Based Production proposes that production is more efficient if speaker mentions material that is readily available first. According to ABP, speaker choice is governed by: 1) when a part of a message needs to be expressed within an utterance; 2) when the linguistic material to encode that part of the message becomes available (K. Bock, 1987; V. S. Ferreira & Dell, 2000). Specifically, if material that encodes a part of the message becomes available when it comes time to convey that part of the message, it will be used. However, if that material is not yet available, then other available material will be used, as long as it is compatible with the grammatical context produced thus far and it does not cut off the speaker's future path to expressing the desired content. This is also referred to as THE PRINCIPLE OF IMMEDIATE MENTION (V. S. Ferreira & Dell, 2000).

Suppose a speaker has just uttered the word *app* in (3) and has in mind to convey the remainder of the utterance meaning as a relative clause. If the word *dancers* becomes available quickly, then according to the principle of immediate mention, a sentence without *that* should be produced (see (3)). If *dancers* does not become available quickly, however, ABP predicts that the speaker will utter *that* to buy more time for *dancers* to become available. (Note that this account relies on an implicit auxiliary assumption that that *that* will generally become available quite quickly; this assumption is rendered plausible by the fact that it is a high-frequency word used in a wide variety of contexts.) If the first word of the RC takes longer to become available the lower its contextual predictability—an assumption consistent with previous work on picture naming (Oldfield & Wingfield, 1965) and word naming (Balota & Chumbley, 1985)—then the less predictable the relative clause, the lower the probability that its first word, *dancers*, will be available at when the speaker reaches the RC, and the higher the probability that the speaker will use *that*. Since an RC is required after *app* in order for it to be followed by the word *dancers*, the lower the contextual probability of an RC the lower the contextual probability of its first word, predicting the empirically observed relationship between phrasal onset probability and optional function word omission rate.

4.2.3 Distinguishing theories of predictability-driven speaker choice

Although UID and ABP are substantially different theories of what drives speaker choice, they make the same prediction for the effect of contextual predictability on optional reduction of function words for cases such as (3). It is thus intrinsically difficult to use optional reduction phenomena to tease these accounts apart. Prior work (Jaeger, 2010) acknowledged this entanglement of the predictions and attempted to tease these accounts apart via joint modeling using logistic regression. There are other phenomena for which the accounts make similar predictions, as well. Consider the case of ordering choices for words or phrases, such as subject–object versus object–subject word order for languages in which both options are available, such as Russian. Availability-Based Production predicts that whichever becomes available earlier will be uttered first (W. Levelt & Maasen, 1981); if the lexical encodings of more contextually predictable references tend to become available more quickly, then more predictable arguments will tend to be uttered first. This prediction is indeed likely to be true: a given-before-new word order preference is widely recognized to influence many languages (Behaghel, 1930; Prince, 1981; Gundel, 1988), and discourse-given entities are generally more contextually predictable than discourse-new entities. But UID turns out to make the same prediction. Two arguments of the same verb generally carry mutual information about each other, so any argument will typically be less surprising if it is the latter of the two. Thus, putting the argument that is more predictable from sentence-external context before the less-predictable argument will lead to a more uniform information density profile and will be preferred.

In the case of speaker choice for Mandarin classifiers, however, UID and ABP turn out to make different predictions as we describe in the next section. The empirical facts regarding speaker choice for classifiers are thus of considerable theoretical interest.

Zhan and Levy (2018) have argued that UID and ABP make different predictions on Mandarin Classifier use with regard to noun predictability. As regards UID, the choice between a specific classifier and a general classifier will typically affect the contextual predictability of the noun modified by the classifier. In particular, a specific classifier constrains the space of possible upcoming nouns more tightly than the general classifier (Klein et al., 2012), thus generally reducing the actual noun's surprisal. The UID hypothesis thus predicts that speakers choose a **specific** classifier more often when the noun predictability would otherwise be low than when the noun is more predictable. This is because the use of a specific classifier makes the distribution of information density more even between the noun and the classifier.

Availability-Based Production, on the other hand, makes different predictions than UID. The fundamental prediction of ABP is that the harder the noun lemma is to access, the less often the speaker will use a specific classifier, provided two plausible assumptions. First, the general classifier *ge* is always available, regardless of the identity of the upcoming noun, as it is the most commonly used classifier and is compatible with practically every noun. Second, in order to access and produce an appropriate specific classifier, a speaker must complete at least some part of the planning process for the production of the nominal reference: accessing the noun lemma, or minimally accessing the key semantic properties of the referent that determine its match with the specific classifier. On these two assumptions, any feature of the language production context that makes the noun lemma less accessible or that more generally makes noun planning more difficult will favor the general classifier. In out-of-linguistic-context picture naming, for example, noun lemma accessibility is known to be driven by noun frequency (Oldfield & Wingfield, 1965). The lower the noun frequency, the less accessible the noun lemma, thus the less likely a specific classifier will be used. To make predictions about the effect of noun predictability on classifier choice in linguistic contexts, we must add a third, theoretically plausible assumption: that less predictable noun lemmas are harder and/or slower to access than more predictable noun lemmas. On these three assumptions, in corpus data the link between noun lemma accessibility and classifier choice will show up as an effect of noun predictability, which by hypothesis is determining noun lemma accessibility. For less predictable nouns, their specific classifiers will less likely be available to the speaker when the time comes to initiate classifier production. Because noun lemmas need to be accessed in order to produce specific classifiers, and the less predictable the noun, the harder the noun lemma is to access and hence the specific classifier associated with the noun becomes available by the time a classifier needs to be produced.

In other words, the link between noun lemma accessibility and classifier choice will manifest in different ways in different tasks, depending on whether one we are looking at usage in linguistic context versus picture-naming. Under our assumptions about ABP, we can identify three predictions. First, in corpus data, speakers should as described above choose a **general** classifier more often the less predictable the noun. Second, in out-of-context picture naming, speakers should as described above choose the **general** classifier more often the more frequent the noun (provided there is high naming agreement for the picture, so that there is not competition among nouns that affects the production process). Finally, we can add a third prediction based on the temporal dependence of specific classifier production on noun planning: when speakers are under greater time pressure, they should produce the general classifier more often, as it can be used even when noun planning has not proceeded far enough for a specific classifier to be available.

Zhan and Levy (2018) tested the first prediction in an investigation of naturally occurring texts, using language models to estimate noun predictability and mixed logistic regression to infer its relationship with classifier choice. They found that the less predictable the noun, the lower the rate of using a specific classifier. While these results lend support for the Availability-Based Production account, the study has some limitations. One limitation is that the corpus being used was a collection of online news texts. Written language may serve as a first approximation of testing theories of language production, but it would be ideal to use real-time language production task to further test the hypotheses. Another limitation is that there was no experimental control of context, so it is possible that predictability was confounded with some other contextual factor that was not included in their regression analysis but that is actually responsible for speaker choice.

In the present study, we use a real-time language production task involving picture naming varying noun frequency and whether speakers are put under time pressure, allowing us to further investigate the two models of language production by testing the first and third predictions described above.

4.3 Experiment 1: Picture-Naming Experiment

We used a picture-naming experiment to test the predictions of Uniform Information Density and Availability-Based Production by manipulating noun frequency and whether or not the speaker is under time pressure. This picture-naming experiment offers a simple yet effective way to elicit real-time language production.

4.3.1 Participants

Fifty-nine self-reported native speakers of Mandarin Chinese were recruited via Witmart, a China-based online crowdsourcing platform. Participants received compensation for their time.

4.3.2 Materials

We adapted images from the Pool of Pairs of Related Objects (POPORO) (Kovalenko, Chaumon, & Busch, 2012) image set to create our visual stimuli. We selected images from the image set based on the following criteria: 1) the image can be described by a count noun; 2) the preferred count noun is compatible with the general classifier and at least one specific classifier. In total, the experiment had 55 items. [2]

[2]The code and materials for constructing the picture-naming experiments are available at `https://github.com/meilinz/liang-zhi`. The processed data and analysis code are available at `https://osf.io/8zwrm/`.

Figure 4-1: Sample visual display for the picture-naming experiment. The red dot below the picture is the recording light. Below it is the text indicating the status of the recorder; in this case, it is "recording stopped". The English translation for the sentence in the bottom is: “Please describe the number and the name of the objects in the picture."

We developed a browser-based version of the experiment using `jsPsych` (de Leeuw, 2015), a JavaScript library for creating behavioral experiments in a web browser. We estimated the frequencies of occurrence of the preferred count nouns from `SogouW` (Sogou, 2006), a word frequency dictionary for online texts in Chinese.

4.3.3 Procedure

After providing consent, participants were presented with scenes of various countable object kinds such as cabbages and tables. Figure 4-1 shows a sample display. In each scene, there were several instances of the same object kind. The number of objects in each trial varied from two to four. Participants were asked to describe the number and the name of the object in Mandarin, eliciting utterances such as “three CL chairs" which we recorded.

Participants were assigned to one of the two conditions. In the **Quick** condition

(N=30), recording started 50 ms after the picture was shown, indicated by a recording light at the bottom of the picture. Each trial ended after 5 seconds of recording, and the next trial began automatically. In the **Slow** condition (N=29), recording stared 3 seconds after the picture was shown, and participants clicked on the screen to move toward the next trial.

4.3.4 Predictions

Availability-Based Production predicts that (1) the rate of specific classifier use will be lower in the **Quick** condition, when speakers are under time pressure, than in the **Slow** condition; and (2) the rate of specific classifier use will be lower for less frequent nouns. This latter prediction derives from evidence that lexical access, as manifested by response latencies, takes longer for lower frequency words in language production experiments requiring word production outside of sentence context; this holds not only for picture naming (Oldfield & Wingfield, 1965), as we require of participants here, but also of visually-presented word naming (Balota & Chumbley, 1985). If lower-frequency nouns are slower to access, their specific classifiers may also be slower to access and thus less often used than the general classifier, which is available for all nouns.

The predictions of Uniform Information Density for the effect of the **Quick**/**Slow** manipulation are unclear. As regards noun frequency, UID predicts that if anything low-frequency nouns should have a *higher* rate of specific classifier usage, as a noun's frequency may effectively serve as its predictability in this experimental setting without broader linguistic context.

4.3.5 Analysis

Audio responses were first transcribed to texts using Google's speech-to-text application programming interface (API), and then checked manually to correct transcription errors. We excluded trials when the participant did not produce a classifier or a noun. For each item, we used the nouns that were most frequently produced as the noun for

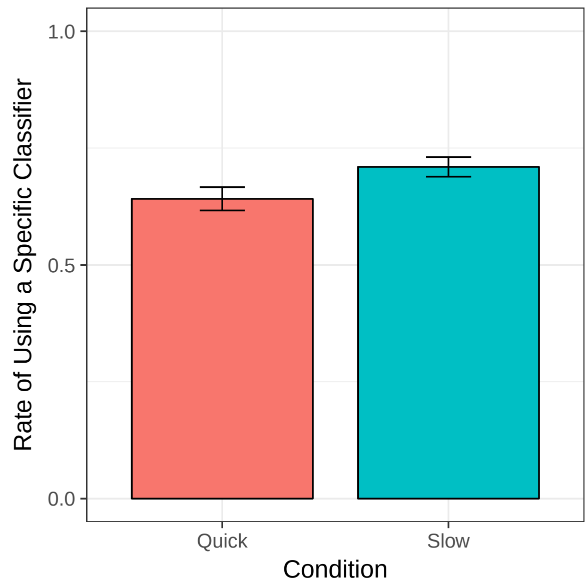

Figure 4-2: Expt 1: Quick vs. Slow manipulation and rate of using a specific classifier as opposed to the general classifier *ge*. Error bars are standard errors over by-participant means.

that item. We also compiled a list of acceptable nouns for each items, and excluded nouns that were not on the list.

We used a mixed-effect logit model to investigate whether noun frequency and time pressure affect classifier choice. The dependent variable was the binary outcome of whether the general classifier or a specific classifier was produced. For each noun type, we also identified its preferred specific classifier (using native speaker introspective judgment and predominant responses by experimental participant, which were concordant). We included two predictors in the analysis: log noun frequency and condition. We included noun, preferred specific classifier, and participant as random factors. We used the maximal random-effects structure with respect to these two predictors (Barr et al., 2013). For condition, this entailed random slopes by noun and by preferred specific classifier, but not participant because the condition manipulation was between subject. For log noun frequency, this entailed a random slope by participant. The full formula in the style of `R`'s `lme4` is:

```
classifier_choice ~ log_noun_frequency + condition
```

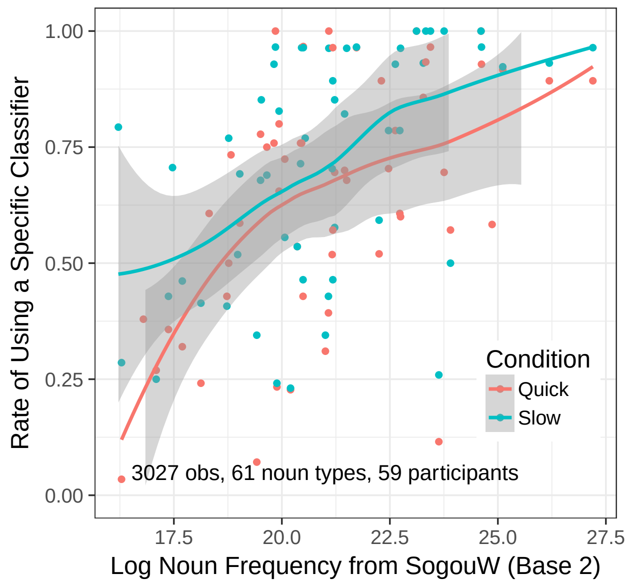

Figure 4-3: Expt 1: The relationship between noun frequency and rate of specific vs. general classifier use in picture naming.

```
+ (1 + condition | noun )
+ (1 + condition | potential_specific_classifier)
+ (1 + log_noun_frequency | participant)
```

Statistical significance was determined using Markov chain Monte Carlo (MCMC) methods in the R package `MCMCglmm` (Hadfield, 2010) with p-values based on the posterior distribution of regression model parameters with an uninformative prior, as is common for MCMC-based mixed model fitting (Baayen et al., 2008).

4.3.6 Results

Looking just at the **Quick**/**Slow** contrast, we find (Figure 4-2) that speakers produced more instances of the general classifier when they are under time pressure than when they are not ($p < 0.05$), suggesting that specific classifiers are slower than the general classifier to access and thus supporting the Availability-Based Production account.

Further breaking out our results by noun log-frequency, we find (Figure 4-3) that the lower frequency the noun, the more likely a **general** classifier is to be produced

($p < 0.001$). This pattern holds within both experimental conditions and is consistent with previous results from the corpus analysis (Zhan & Levy, 2018, Chapter 3 of this thesis), and also supports the Availability-Based Production account.

One potential concern arises in the frequencies of the different specific classifiers. One could argue that it was not the noun's frequency that determined the use of the general classifier, rather it was the frequency of the preferred specific classifier that affected the choice of which classifier was used. In the mixed-effect logit model presented above, we included a by-specific-classifier random intercept, which largely rules out the possibility that specific classifier frequency were confounding the effect of noun frequency. To further investigate this issue, we tried a version of our regression model that also includes a fixed effect for the log frequency of preferred specific classifier as a control factor. We did not find any qualitative change to the results. The effects of noun frequency ($p < 0.001$) and condition ($p < 0.05$) on classifier choice remain qualitatively similar to the results of the original model. Furthermore, in this new analysis, there is no effect of specific classifier frequency on classifier choice ($p = 0.483$). This additional analysis suggests that it is unlikely that specific classifier frequency to be driving the effect of noun frequency.

4.3.7 Discussion

Using a picture-naming experiment, we show that Availability-Based Production predicts speakers' real-time choices of Mandarin Chinese. The lower a noun's frequency, the more likely a general classifier is to be used. We also found that the use of classifier is moderated by whether the speaker is under time pressure when speaking, where the speaker tends to produce more instances of the general classifier if they are under greater time pressure to speak. This real-time effect confirms that the general classifier is easily accessible when the speaker is about to produce a noun phrase with numeral.

Although we have not found direct evidence supporting the UID hypothesis, it is possible that this particular experimental setting is not very communicative in nature. In Experiment 2, we performed real-time language production experiment in a more communicative setting, with a human listener present in the experiment to further test speaker choice in language production.

4.4 Experiment 2: Reference Game Experiment

Experiment 2 is very similar to Experiment 1 with addition of a listener role to create a reference game. We used the same materials from Experiment 1. In addition, we added 9 new items that also satisfy the same criteria specified in Experiment 1. In total, the experiment had 64 items. We also created corresponding materials for the listener player in this game.

4.4.1 Participants

Seventy self-reported adult native speakers of Mandarin Chinese from Massachusetts Institute Technology and nearby communities participated in the experiment. Participants received compensation for their time.

4.4.2 Procedure

There were two players in the reference game: a speaker (the participant), and a listener (one of our Chinese-speaking research assistants). Participants were randomly assigned to one of the eight lists. Each list had 4 blocks, and each block consisted of 16 items, presented in either the Quick condition, or the Slow condition. After providing consent and filling out a linguistic background questionnaire, the experimenter (the first author or another Chinese-speaking research assistant) introduced the listener player to the participant. In the experiment, the participant (the speaker) sat face to face on a table with a listener in this game. They both had a laptop in front of them, but they could not see each other's screen.

From the speaker's side, participants saw the same browser interface as the one used in Experiment 1 (see Figure 4-1). Participants were told to describe the number and the name of the objects displayed in a phrase in Mandarin so that the listener

Figure 4-4: Sample visual display for the listener in the reference game experiment.

can pick up the right image from a grid of four images, as shown in Figure 4-4.

If the listener clicked on the right image, they would hear a clear DING sound indicating success. If the listener clicked on the wrong image, they would hear a BOMB sound indicating failure of communication. If not enough information was given (e.g., the speaker did not say anything or said something intelligible), the listener would click on a button that also produced a BOMB sound.

From the listener's side, the research assistant saw a grid of four images and only one of them corresponding to the one on the speaker's side. The listener was blind to which list the speaker was on, so they did not know the right answer for each trial. They were instructed to click on the corresponding image or button based on the speaker's utterance. Both the speaker and the listener only knew whether the trial was a success at the time the listener responded in clicking.

4.4.3 Analysis

We used the same pipeline from Experiment 1 to process and analyze the audio data. Because Condition is within-subject in the experiment, the statistical analysis is slightly modified to account for the change in the experimental design (adding a random slope for Condition in the participant random effect). The full formula in the style of `R`'s `lme4` is:

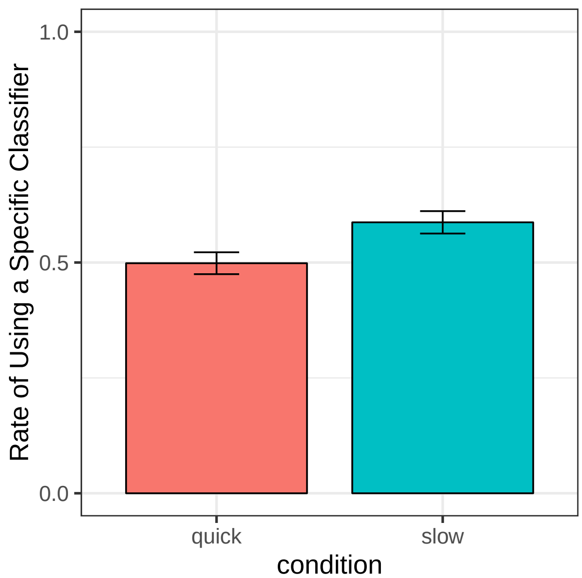

Figure 4-5: Expt 2: Quick vs. Slow manipulation and rate of using a specific classifier as opposed to the general classifier *ge*. Error bars are standard errors over by-participant means.

```
classifier_choice ~ log_noun_frequency + condition
+ (1 + condition | noun )
+ (1 + condition | potential_specific_classifier)
+ (1 + log_noun_frequency + condition| participant)
```

4.4.4 Results

The results from Experiment 2 largely replicated what we found in Experiment 1 qualitatively. There was a main effect of condition (Quick vs. Slow) ($p < 0.05$) as shown in Figure 4-5, where speakers were more likely to produce the general classifier under greater time pressure. Notice that this time the time pressure manipulation was within-subject (Condition was between-subject in Experiment 1), which provides stronger evidence for the time pressure effects. There was a main effect of log noun frequency ($p < 0.001$) as shown in Figure 4-6, where speakers were more likely to produce a specific classifier for more frequent nouns.

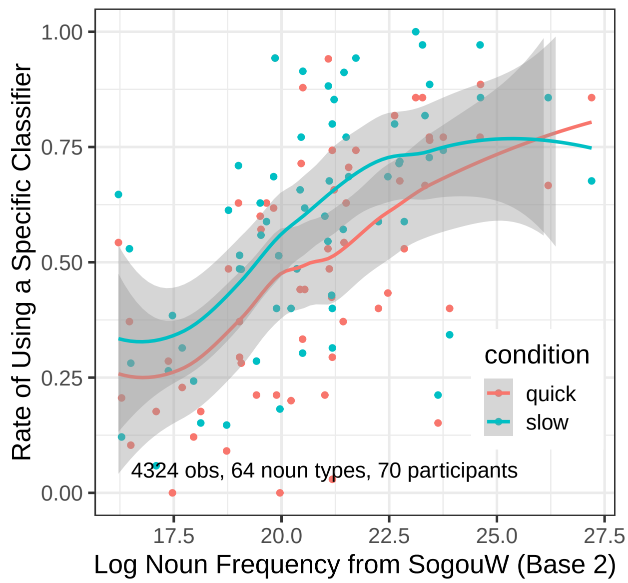

Figure 4-6: Expt 2: The relationship between noun frequency and rate of specific vs. general classifier use in picture naming.

4.4.5 Discussion

Experiment 2 was designed as a reference game aiming to incorporate a communicative element to the real-time language production task. The speaker's communicative goal was to say something to help the listener pick out the right picture. The results from this experiment confirmed two predictions by Availability-Based Production. First, speakers are more likely to use the general classifier under greater time pressure to produce. Second, the rate of using a specific classifier is lower with less frequent nouns. These results provide further support for a speaker-oriented model of speaker choice in language production, indicating that speaker choice is, at least partially, a by-product of speaker-centric information processing.

Overall, there is a drop in the rate of using a specific classifier in Experiment 2 compared to the results in Experiment 1. This difference may be related to the following three changes between the two experiments:

Participants' linguistic background Participants in Experiment 2 were recruited in the United States. Although all of them were native speakers of Chinese and all of them speak fluent Mandarin, they might have various degree of first language attrition. Participants in Experiment 1 were native Chinese speakers based in China. Chinese is likely to be their primary and if not the only language they use on a daily basis. This might be the main reason for the overall decrease in specific classifier use in Experiment 2.

Between-subject vs. within-subject design Experiment 2 used a within-subject block design to expose the participants to both the Quick and the Slow conditions, whereas in Experiment 1, each participant was randomly assigned to one of the two conditions. It is possible that the within-subject design made it harder for the participant to be consistently using a specific classifier, as half of the time they were under greater time pressure to speak.

The presence of a listener It is possible that the presence of a listener may contribute to the change of classifier use in Experiment 2. It could be that the speaker was more likely to be nervous in front a human listener, and that affects their performance in speaking. But this could easily go the other way. The speaker might want to impress the listener and would try harder to use a specific classifier when possible. We had little insights to pin it down based on the data we currently have.

Despite the quantitative differences between Experiment 1 and Experiment 2, we still see the robust effects of noun frequency showing in Experiment 2. The effect of speed (time pressure) manipulation also replicated in Experiment 2.

4.5 Conclusion

Using a picture-naming experiment and a reference game experiment, we show that Availability-Based Production predicts speakers' real-time choices of Mandarin Chinese. The lower a noun's frequency, the more likely a general classifier is to be used. We also found that the use of classifier is moderated by whether the speaker is under time pressure when speaking, where the speaker tends to produce more instances of the general classifier if they are under greater time pressure to speak. This real-time effect confirms that the general classifier is easily accessible when the speaker is about to produce a noun phrase with numeral. The effects of time pressure and noun frequency remain qualitatively similar regardless of whether a listener is present in the experiment.

As we have illustrated in Section 4.2.3, the link between noun lemma accessibility and classifier choice would manifest in different ways depending on different tasks. For classifier usage in rich linguistic context such as the corpus study in Chapter 3, noun surprisal serves as a good proxy for estimating predictability, and the effects of noun frequency is regressed out. For out-of-context picture naming, noun frequency effectively serves as an indicator or predictability, and lemma accessibility more generally.

Taken together, the present study and previous corpus work on Mandarin classifier (Zhan & Levy, 2018) offer converging evidence regarding the relationship between noun frequency, predictability, and classifier choice, and thus shed light on the mechanisms influencing speaker choice. While the corpus work provides ecological validity through naturalistic data, the experimental work helps us to eliminate potential correlation-based confounds with a clean setup, and enables us to get dense data that are theoretically important but naturalistically sparse. When combined together, this work is complementary with previous corpus work and together paint a more comprehensive picture of language production.

Chapter 5

Effects of Counting on Mandarin Classifier Production

5.1 Introduction

In English and other languages with a grammaticalized count–mass distinction, count nouns such as *chair* can be used with a numeral directly and typically exhibit singular–plural morphology marking (e.g., one chair, three chairs). In Mandarin and other classifier languages, nouns lack obligatory singular–plural marking and a numeral cannot be used a with noun directly. A numeral classifier is required to specify the quantity (e.g., one CL chair, three CL chair). While different nouns are compatible with different SPECIFIC classifiers, there is a GENERAL classifier *ge* (CL.general) that can be used with most nouns. Often, the use of general versus specific classifier for a given noun carries little to no meaning distinction for the utterance, as illustrated (1) and (2) below.

(1) 我 买了 一台 照相机
wo mai-le yi **tai** zhaoxiangji
I bought one CL.machinery camera (“I bought a camera")

(2) 我 买了 一个 照相机
wo mai-le yi **ge** zhaoxiangji

I bought one CL.general camera (“I bought a camera")

In this study, we focus on speaker choice between the general classifier and specific classifier where both options express the same or very similar meaning. The use of specific classifier is found to occur more often with: 1) nouns that are more predictable in context in the corpora (Zhan & Levy, 2018); 2) higher-frequency nouns in real-time language production such as picture-naming (Zhan & Levy, 2019). Zhan and Levy (2019) also found that the use of classifier is moderated by whether the speaker is under time pressure when speaking, where the speaker tends to produce fewer instances of specific classifiers if they are under greater time pressure to speak.

The results of Chapters 3 and 4 (Zhan & Levy, 2018, 2019) are consistent with the hypothesis that producing a specific classifier requires deployment of linguistic knowledge (i.e. an appropriate specific classifier) that is accessed through the noun lemma and takes time and/or cognitive resources: if we give the speaker less time and/or make the noun lemma more taxing to retrieve or produce, we get fewer specific classifiers. One working assumption from the two studies is that to access and produce a specific classifier, speakers must complete at least some part of the planning process for the production of the nominal reference: accessing the noun lemma, or minimally accessing the key semantics properties of the referent that determine its match with the specific classifier. Is this a reasonable assumption to make? In this chapter, we use a counting task to probe the lexical access mechanism during classifier production.

Adults estimate small numbers (1-4 or sometimes up to 5) rapidly and accurately (Kaufman, Lord, Reese, & Volkmann, 1949; Mandler & Shebo, 1982), often referred to as subitizing (Kaufman et al., 1949). Previous studies (Zhan & Levy, 2018, 2019) only looked at classifier production with small number of objects (up to 4). In those settings, counting was most likely not required. What is lexical access like in a classifier language when counting is required to express the exact quantity of objects? Does counting multiple identical instances of objects enhance or inhibit the access and production of relevant specific classifiers?

Previous work argues that there are three stages or levels of processing in the time course of lexical access, in particular in the process of naming of objects (W. J. Levelt, 1989, Chapter 6). The first stage is the visual processing of the picture. This stage is sensitive to interference from visually similar materials. The second stage is conceptual categorization, where preference for basic object level (e.g., apple, chair, dog) is found (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). The third stage is lexical access proper, which is sensitive to word frequency effects (Oldfield & Wingfield, 1965). Incorporating counting to classifier production provides an opportunity to investigate the nature of semantic access that required to get a specific classifier. During picture-naming, to produce a noun phrase with a classifier and a quantity that requires them to count, at least the following three components are needed: counting, retrieving a noun, selecting a classifier. Which stage of processing in the time course of lexical access does these three components belong to? Does classifier selection complete after the noun retrieval? Does classifier selection complete in the second stage (i.e., the stage of conceptual categorization), or in the third stage (i.e., the stage of lexical access proper)? By manipulating the number of objects, we could use counting to help us investigate these questions.

It has been observed that speakers sometimes speak and plan simultaneously with incremental production (W. J. Levelt, 1989; V. S. Ferreira, 1996; F. Ferreira & Swets, 2002; Brown-Schmidt & Konopka, 2015). When having people do arithmetic problems while saying the answer out loud in a complete sentence (e.g., The answer is 32), both latencies and utterance durations are longer for harder problems (F. Ferreira & Swets, 2002), indicating language production is incremental but not entirely incremental as it could be. The longer latencies before articulating the answer for harder problem suggests that speakers do think and plan the answer before they initiate the sentence, and the longer utterance duration indicates speakers also take advantage in the process of producing an utterance so that can speak and plan at the same time.

In the case of classifier production with counting, it could be that, during counting of multiple instances of the same object, the noun (and its relevant semantic and syntactic information including possible specific classifiers) is accessed repeatedly or in higher activation level; this would predict that objects in quantities large enough to require counting would yield a higher rate of specific classifiers, relative to objects within subitizing quantities. Alternatively, it could be that counting of multiple instances does not involve access to the noun (e.g. the number of objects could be ascertained before the object category is even fully determined or before the noun form is fully retrieved). In that case, one would get a lower rate of specific classifiers because counting leaves the speaker with less time/resources for noun lemma and then specific classifier access than subitizing does.

Another possibility in the case where bigger numbers decrease the rate of using a specific classifier could simply be that smaller numbers are more frequent than bigger numbers. There are a lot of instances of the form "two CL Noun", but fewer instances of the form "eleven CL Noun". If everything else stays constant, smaller numbers get more exposure to specific classifiers due to the frequency advantage. If there is an effect due to number manipulation, we will check the corpus data to understand whether it is the effect of number (or numeral–classifier frequency), or the effect of counting. If higher numeral–classifier frequency co-occurs with higher rate of using a specific classifier, then the results could be a by-product from the imbalanced distribution of numeral–classifier frequencies in the language. If higher numeral–classifier frequency does not co-occur with higher rate of using a specific classifier, then the results are likely about counting (counting takes cognitive resources that would otherwise facilitate classifier availability), not just about number or numeral–classifier frequency *per se*.

5.2 Hypotheses and predicted outcomes of classifier production

From the previous session, we have the following two opposing predictions: 1) If the speaker has to count, this gives them more time to access the noun lemma, and also more time looking at the objects, assuming that specific classifiers are associated with noun lemma, the speaker will be more likely to produce a specific classifier. 2) If the speaker has to count, this taxes their cognitive capacity and therefore, they will be more likely to go with the general classifier.

To pin down the above ideas and taking the time course into consideration, below I illustrate four hypotheses and their predicted outcomes on classifier choice. I use an arrow to indicate the procession in time, but it does not necessary need to be strictly sequential in the specified order. It is possible that some stages could start while the previous one is still ongoing, as long as the previous stage starts before the current stage. Speaking can happen anytime after counting is finished.

Hypothesis 1: Count objects & Select CL → Retrieve Noun The speaker counts the objects and select the classifier at the same time, and counting facilitates classifier retrieval at the conceptual categorization level, before the noun is accessed. This predicts that greater number leads to more specific classifiers because counting activates access to potential specific classifiers associated with the objects.

Hypothesis 2: Count objects → Retrieve Noun → Select CL The speaker counts the objects, retrieve the noun, and select the classifier to use. This predicts that greater number leads to fewer specific classifier, because counting takes up time and resources for noun access and access to specific classifiers. The speaker would be more likely to use the general classifier when counting is needed.[1]

Hypothesis 3: Count objects & Retrieve Noun → Select CL The speaker counts objects and retrieve the noun at the same time, and that counting facilitates noun retrieval by looking at multiple instances of the objects. This predicts that greater number leads to more specific classifier because counting extends the time for

[1]One potential caveat is that how can we count before we have done enough object individualization and conceptualization before we know what we are counting? To address this, we might want to think about the second stage "Retrieve Noun" as retrieving the noun form. It is possible that one already access the concept of the object associated with the noun, but has not completed access the form yet (e.g., Tip of the Tongue effect (Brown & McNeill, 1966)). The same applies for classifier. However, this would not change the predicted outcomes of this hypothesis.

accessing the noun and its associated specific classifiers. This hypothesis makes the same predictions as Hypothesis 1, but for different reasons.

Hypothesis 4: Retrieve Noun → Select CL → Count objects One could retrieve the noun, select the classifier to use, and then count the objects. This predicts there will be no effects of increasing the number beyond subitizing range (counting is required) on classifier choice, because the noun is accessed and the classifier is selected before counting. This is not the best strategy, at least not for a language in Mandarin Chinese, where the word order is Numeral Classifier Noun. Since the numeral needs to be said first, the speaker would need to hold on to the noun and the classifier. They would not be able to start the phrase properly until they finish counting.

5.3 Experiment 3: Counting Game Experiment

We adapted the picture-naming paradigm from Experiments 1 & 2 (Zhan & Levy, 2019) to a Counting Game Experiment, where we increase the number of objects beyond the subitizing range so that participants had to count to get the exact number.

5.3.1 Participants

Fifty-six self-reported native speakers of Mandarin Chinese were recruited from the MIT community. One participant were excluded for his overall accuracy were not above 75%. Participants received compensation for their time.

5.3.2 Materials

We adapted images from the Pool of Pairs of Related Objects (POPORO) (Kovalenko et al., 2012) to create the stimuli. Following recent work on classifier production (Zhan & Levy, 2019), we selected images from the image set that satisfy the following criteria: First, the image can be described using a count noun; Second, the preferred count noun can be paired with either the general classifier or at least one specific

Figure 5-1: Sample visual display for the picture-naming experiment (Small number).

Figure 5-2: Sample visual display for the picture-naming experiment (Big number).

classifier. In total, the experiment has 64 items. [2]

The experiment has a 2 by 2 design with items varying in their noun frequency. In out-of-context picture naming, noun lemma accessibility is known to be driven by noun frequency (Oldfield & Wingfield, 1965). We estimated noun frequency from SogouW (Sogou, 2006), a word frequency dictionary of online texts in Mandarin Chinese. The two factors are Speed (Fast vs. Slow, i.e., whether the speaker has to start speaking as fast as possible) and Number (Small number ($2-5$), Big number ($7-11$)). Both Speed and Number are within-item and within-subject, and noun frequency is between-item and within-subject. Each item has four variants in terms of number: half of them are Small, and half of them are Big.

[2]The data and analysis code are available at `https://osf.io/8zwrm/`.

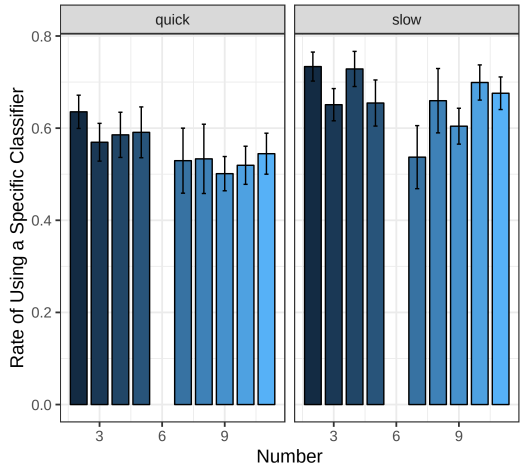

Figure 5-3: Expt 3: Number manipulation and rate of using a specific classifier as opposed to the general classifier *ge*. Error bars are standard errors over by-participant means.

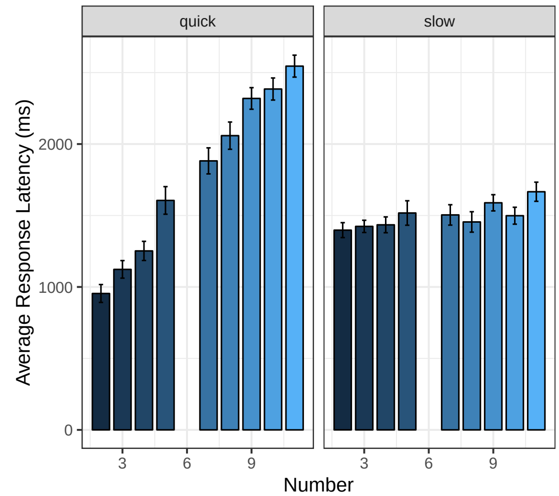

Figure 5-4: Expt 3: Response latency in milliseconds against number of objects. Error bars are standard errors over by-participant means.

5.3.3 Procedure

Participants were randomly assigned to one of the 16 lists. Each list has 4 blocks, and each block consists of 16 items, presented in either the Quick condition, or the Slow condition. Participants were presented with scenes of various countable object kinds such as guitars and chairs. Each scene has several identical instances of the same object kind. The number of objects in each trial varied. Participants were asked to describe the number and name of the object in Mandarin, eliciting utterances such as "three CL guitars". Their responses were recorded.

Following Zhan and Levy (2019), in the **Quick** condition, recording started 50 ms after the display of the objects, indicated by the recording light at the bottom of the picture. Each trial ended after 5 seconds of recording, and progress to the next item began automatically. In the **Slow** condition, recording started 4 seconds after the picture was shown, and participants clicked on the screen to proceed to the next item. Half of the trials presented objects in **Small** numbers (2-5), and the other half in **Big** numbers (7-11). The order of presentation within each block was randomized.

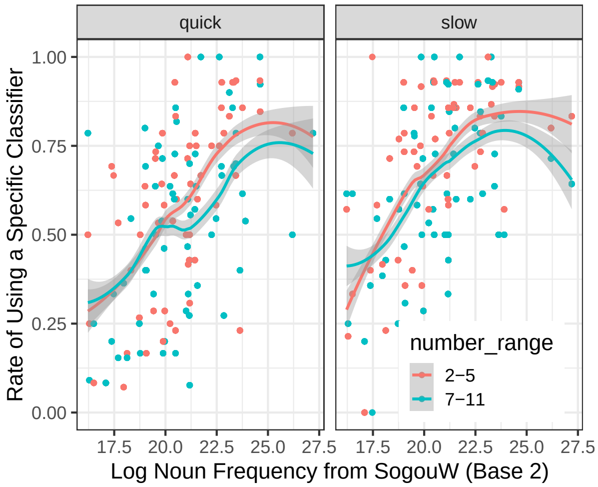

Figure 5-5: Expt 3: Noun frequency, Number, and rate of specific vs. general classifier use.

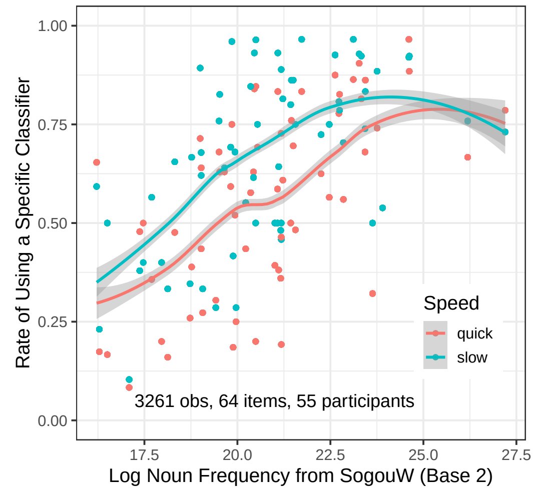

Figure 5-6: Expt 3: Noun frequency, Speed, and rate of specific vs. general classifier use.

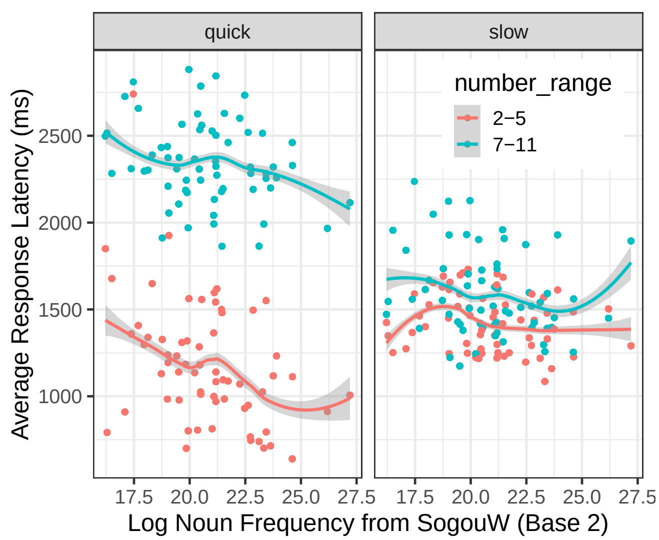

Figure 5-7: Expt 3: Response latency in milliseconds against noun frequency

5.3.4 Analysis

Audio responses were first transcribed to texts using Google's speech-to-text application programming interface (API), and then checked manually to correct transcription errors. We excluded trials when the participant did not produce a classifier or a noun, or the number of objects was wrong. For each item, we used the nouns that were most frequently produced as the noun for that item.

We used a mixed-effect logistic model to investigate whether noun frequency, time pressure, and number affect classifier choice. The dependent variable was the binary outcome of whether the general classifier or a specific classifier was produced. For each noun type, we also identified its preferred specific classifier (using native speaker introspective judgment and predominant responses by experimental participant, which were concordant). We included three predictors in the analysis: log noun frequency, speed, and number. We included noun, preferred specific classifier, and participant as random factors. We used the maximal random-effects structure with respect to these two predictors (Barr et al., 2013). For speed, this entailed random slopes by participant, by noun and by preferred specific classifier. For log noun frequency, this entailed a random slope by participant. For number, we included random slope by item, by participant, by preferred specific classifier. The full formula in the style of `R`'s `lme4` is:

```
response ~ log_noun_frequency + speed + number
+ (1 + speed + number | noun )
+ (1 + speed + number | preferred_specific_classifier)
+ (1 + log_noun_frequency + speed + number| participant)
```

Statistical significance was determined using Markov chain Monte Carlo (MCMC) methods in the R package `MCMCglmm` (Hadfield, 2010) with p-values based on the posterior distribution of regression model parameters with an uninformative prior, as is common for MCMC-based mixed model fitting (Baayen et al., 2008).

5.4 Results

5.4.1 The use of classifiers

As illustrated in Figure 5-3, there was a main effect of Speed ($p < 0.001$), where the rate of using a specific classifier is lower in the Quick condition than in the Slow condition, consistent with the findings by Zhan and Levy (2019). In addition, there was a main effect of Number ($p < 0.05$), where the bigger the number, the fewer specific classifiers were used. The results are aligned with Hypothesis 2, and at the same time, excluded the other three hypotheses listed in Session 5.2. In particular, the results has ruled out Hypothesis 1 and is compatible with only Hypothesis 2, it is safe to say that the working assumption in Chapters 3 and 4 is reasonable: to produce a specific classifier, speakers need to complete at least part of the planning process for the production of the nominal reference by accessing the noun lemma. The results also rule out Hypothesis 3 and Hypothesis 4, which indicates that it is hard to count and retrieve noun at the same time, and counting is likely to happen first before noun retrieval in classifier production.

Further breaking out our results by noun log-frequency, as shown in Figures 5-5 and 5-6 we find that the higher frequency the noun, the more likely a specific classifier is to be produced ($p < 0.001$). This pattern holds within both experimental conditions and is consistent with previous results from the corpus analysis (Zhan & Levy, 2018) and picture-naming experiment (Zhan & Levy, 2019). Note that here the effects of noun frequency on classifier choice also supports the idea that accessing the noun lemma is likely to happen before selecting a specific classifier. Lower frequency nouns had weaker accessibility and this led to fewer specific classifiers being produced. If it is the other way around (e.g., Hypothesis 1), we probably would not see the effects of noun frequency on classifier production. In short, the data observed in this experiment supports Hypothesis 2.

5.4.2 Response latency

It is plausible that higher numbers force participants to spend longer preparing their utterance, and that retrieval of the classifier cannot easily happen while they are counting. This would give participants effectively less time in high-number conditions (regardless of whether it's the quick or slow condition), and affecting classifier choice in similar ways as the quick vs slow manipulation does.

Further analysis on response latency (Figure 5-4) shows that the average response latency increases as the number of objects increases. Here we define Response Latency as the amount of time in milliseconds it took for participants to utter the first word once the recording started. For the Slow condition, the recording did not start until 4 seconds after the picture was shown. The average response latency in the Slow condition did not change much as the number of objects increases and it is likely that participants had relatively sufficient time to prepare to speak in those extra 4 seconds (not included in the response latency) and hit the ceiling effect.

We performed a multiple regression analysis to analyze the potential effects of Number, Noun Frequency, and Speed on Response Latency. There was a main effect of Number, where bigger number required longer response latency ($p < 0.001$). There was a main effect of Speed, where the Slow condition had a smaller response latency overall. As illustrated in Figure 5-7, there was a main effect of Noun Frequency, where higher frequency nouns had a shorter response latency on average ($p < 0.001$), and this effect is magnified in the Quick condition where participants were under greater time pressure to speak.

The results from response latency alone are not sufficient to distinguish the four hypotheses we have listed in Section 5.2. However, considering the results from classifier choice and response latency, counting of multiple instances does not necessarily involve access to the noun lemma and/or facilitate retrieval of specific classifiers. Having to count gave participants less time and led to more use of the general classifier and longer response time. The overall results support Hypothesis 2.

5.4.3 Additional analysis on corpus data

The results from Experiment 3 found that smaller number within the subitizing range elicited more specific classifiers compared to relatively larger numbers that requires counting during real-time production. One possible reason could be the potential effects of numeral–classifier frequency: simply smaller numbers are more frequent than big ones. For example, there are a lot of instances of the form "two CL Noun", but much fewer instances of the form "eleven CL Noun". If everything else being the same, smaller numbers are more frequently used with a specific classifier due the the frequency advantage. In other words, is the effect of number on classifier choice a by-product from the imbalanced distribution of numeral–classifier frequencies in the language?

If higher numeral–classifier frequency co-occurs with higher rate of specific classifier use, then the results of number manipulation in the experiment could be a frequency-driven by-product due to the fact that smaller numbers are more frequently used with a (any) classifier. If higher numeral–classifier frequency does not co-occur with higher rate of specific classifier use, then the results likely speaks to the effect of counting: counting takes cognitive resources that would otherwise facilitate classifier availability.

Corpus To answer this question, we looked at corpus data and performed additional analysis. We used the dev set from Chapter 3 (Zhan & Levy, 2018) (121,826 observations) with raw data originated from part of SogouCS (Sogou, 2008), a collection of online news texts. We extracted 86,011 classifier-noun pairs with numerals from 1 to 12, accounting for 70.6% of the dev set. We calculated the numeral–classifier frequency based on the entire processed Sogou CS data set (2,197,354 observations). The numeral–classifier frequency is defined as the total counts of numeral-CL in the corpus. For example, the total counts of "five/5 CL" appeared in the corpus. We use numeral–classifier frequency instead of Number frequency because numerals can be used without classifiers in a variety of contexts. We also checked the correlation between Number frequency and numeral–classifier frequency in the corpus, and they

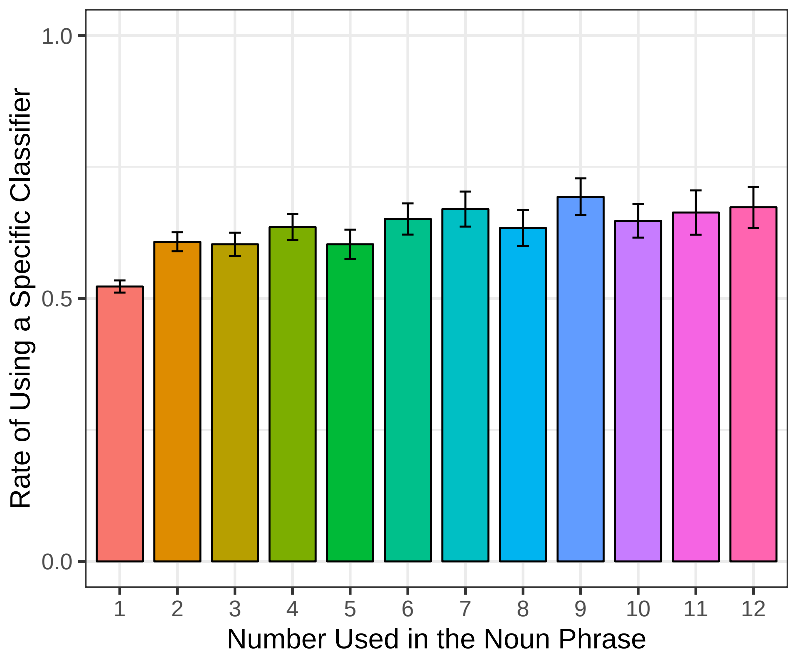

Figure 5-8: Classifier choice and number (1-12) used in the noun phrase from corpus data

are highly correlated ($r = 0.991, p < 0.001, n = 12$). In the following analysis, we will stick to using numeral–classifier frequency (log-transformed) [3].

Analysis We adopted a mixed-effect logistic regression from Chapter 3 (Zhan & Levy, 2018), adding Number and log numeral–classifier frequency to the original analysis. The full formula in the style of `R`'s `lme4` package is:

```
outcome ~ noun_surprisal + log_noun_frequency + number + log_num_cl_freq
+ (1 + noun_surprisal + number + log_num_cl_freq | noun )
+ (1 + noun_surprisal | preferred_specific_classifier)
```

Figure 5-8 shows the number used in the noun phrase (1-12) against the rate of using a specific classifier. To better visualize numeral–classifier frequency with numeral, Figure 5-9 shows the relationship between log numeral–classifier frequency and rate of using a specific classifier, with the numeral labelled under each data point. There was

[3] When replacing numeral–classifier frequency with number frequency, the following analysis had qualitatively the same results.

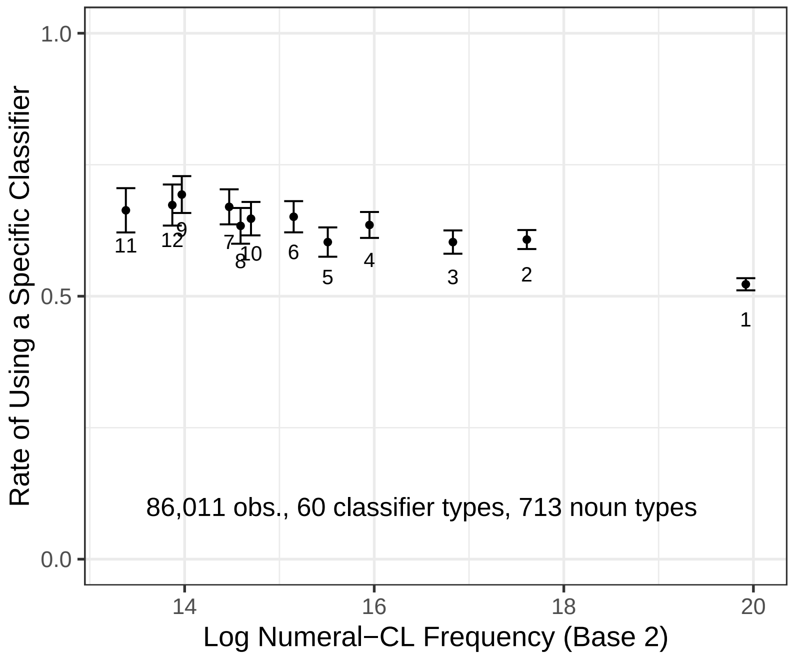

Figure 5-9: Classifier choice and number-classifier frequency from corpus data

no effect of Number ($p = 0.769$). However, there was a main effect of log numeral–classifier frequency, where higher numeral–classifier frequency co-occurred with fewer specific classifiers ($p < 0.001$). From Figure 5-9, we can see that smaller numbers tend to be more frequent than bigger numbers. There is a negative correlation between Number and numeral–classifier frequency ($r = -0.611, p = 0.035, n = 12$).

In addition, there was a main effect of noun surprisal ($p < 0.001$), where nouns with higher surprisal get fewer specific classifiers. There was no effect of noun frequency ($p = 0.794$). These are consistent with the results from the original analysis in Chapter 3 (Zhan & Levy, 2018).

We did not find the effect of number on classifier choice from corpus data[4]. Additionally, the effect of numeral–classifier frequency found in corpus seems to be different than a typical frequency-driven account (e.g., the one in the earlier part of this sub-

[4]When the number 1 was excluded from the analysis, the effect of numeral–classifier frequency went away ($p = 0.215$), and the effect of number became significant ($p = 0.007$). Everything else is qualitatively the same with the original analysis reported in this section. However, the direction of the number effect is that bigger numbers co-occur with more specific classifier use. This is different than a typical frequency-driven account and different than the results from the counting experiment.

section). This indicates that the effect of number from the counting game is unlikely due to the frequency advantage brought by smaller numbers. To be more precise, it is not the effect of Number *per se*, but rather, the effect of Counting during language production.

I think there are a couple of possible reasons why we see different patterns in the corpus compared to the counting game experiment. First, the corpus data is from a collection of online news texts, and in general real time counting behavior is not likely to be happening there. Second, there might be something special going on with "yi-ge" (one CL.general), which could be used to refer to things in a more general sense, similar to "a Noun" in English. This might be part of the reason why we see fewer specific classifiers used with the numeral one in the corpus. We can see that from the corpus data, the numeral one has the highest frequency and the rate of using a specific classifier is the lowest. Note that in the counting experiment and the other experiments reported in the thesis, we did not include the numeral one in the items.

5.5 Conclusion

In this study, we replicated the effect of noun frequency and speed on classifier production found by Zhan and Levy (2019). Namely, the higher the noun frequency, the more likely a specific classifier is to be used. The greater the time pressure, the less likely a specific classifier is to be used, supporting Availability-based Production.

On top of that, we found a novel effect of counting on classifier use. When the number of objects is beyond the subitizing range, the rate of using a specific classifier dropped and response latency increased. This indicates that higher numbers force speakers to spend longer time preparing their utterance, and classifier retrieval cannot easily happen while they are counting.

Our results have ruled out three hypotheses, including one where classifier is selected before accessing to the noun lemma (Hypothesis 1). The data indicate that producing a specific classifier requires deployment of linguistic knowledge that is accessed through the noun lemma and takes time and/or cognitive resources. Our findings support the hypothesis that when the speaker has to count, this taxes their cognitive capacity and therefore, they will be more likely to go with the general classifier. From the results, it seems that the speaker counts the objects, retrieves the noun, and then selects the classifier to use (Hypothesis 2). Based on the effect of counting observed in this experiments and the effect of noun frequency seen in all three experiments reported in the thesis, it seems that selecting a specific classifier requires noun retrieval, and is likely to complete at the stage of lexical access proper.

One might wonder how can one count before retrieving the noun since one probably needs object individualization and conceptualization for counting objects. It is possible that one already access the concept of the object associated with the noun, but has not completely accessed the form yet. Our data have pushed us to the edge of knowledge in this case as this study opens up more new questions for future research. Coupled with our latency response analysis, this study paints a more comprehensive picture of the effects of noun frequency, time pressure to speak, and whether counting is required on speaker choice and response latency in classifier production. Additional analysis of the corpus data also revealed that the effects of Number on classifier production in the counting game are likely not about the number *per se*. Rather, the effects are likely results of counting during language production.

It is intriguing to see that the effect of noun frequency on response latency is smaller than the effect of counting, though the effect of noun frequency on specific classifier rate is much larger than the effect of counting. Perhaps linguistic factors carrying information about language experience such as noun frequency play an important role in speaker choice, whereas cognitive factors such as having to count (especially under time pressure) have a bigger effect on response latency and a smaller effect on speaker choice.

This study also opens up some new questions for future research in cognitive science. For example, what effect would manipulating the ease of visual apprehension of the object category have on the relationship with number? If we had people do some other task such as remember a set of digits or words, would we get the same effect on classifier production?

Viewed more broadly, using classifier production as a test case has helped us investigate whether and how various linguistic and cognitive factors affect speaker choice in real-time language production, and advance our understanding in lexical access and language production more generally.

Chapter 6

Conclusion

This work investigates theories of language production: what general principles govern speaker choice when multiple options are available to express more or less the same thing? This is part of a set of larger questions in language – Why language is the way it is? How do we account for optionality and variation in language production?

To answer these questions, this study combines corpus analysis and psycholinguistic behavioral experimentation using techniques from statistics, natural language processing, and experimental psychology. The corpus component means that this work can detect patterns in naturalistic data and provides ecological validity. The controlled laboratory experiments help eliminate potential correlation-based confounds, and enable us to get dense data that are theoretically important but naturalistically sparse. In addition, this work enriches research in classifier languages by investigating theory-driven question in real-world datasets and controlled experiments.

Both the corpus work (Chapter 3) and controlled experiments (Chapter 4) show that nouns with low predictability are more likely to be paired with a general classifier than nouns with high predictability (or high noun frequency in out-of-context picture-naming), supporting a speaker-oriented availability-based model of speaker choice in language production. The real-time language production experiments also reveal effects of time pressure imposed on the speaker: speakers produced more instances of the general classifier when they were under time pressure to speak.

In a follow-up study (Chapter 5), I used the experimental paradigm for studying classifier production to test whether there are effects of counting (as opposed to subitizing) on classifier choice. Namely, does counting enhance or prohibit the use of a specific classifier in picture naming? This follow-up study demonstrates the possibility of using the experiment paradigm for studying classifier production to address new questions in cognitive psychology and language.

To conclude, I will expand on some theoretical ideas relevant to the work presented in the content chapters, and give further speculations on theoretical implications of the findings.

6.1 Availability-based speaker choice

Using corpus analysis and real-time language production experiments, we found that speaker choice is made to permit early mention of available material. The availability-based effects show up in areas related to **linguistic experience** such as contextual predictability and word frequency, as well as **real-time production or cognitive pressures** such as speaking under time-pressure, and producing an utterance while counting is required.

We found that predictable nouns in the context and frequent nouns in out-of-context picture-naming are more likely to occur with specific classifiers. With predictable or frequent nouns, their associated specific classifiers are more likely to become available at the time of production and thus to be produced. With unpredictable or rare nouns, their associated specific classifiers are less likely to become available at the time of production, leading the speaker to use the easily available default option: the general classifier.

When put under greater time-pressure to speak, the rate of using a specific classifier dropped significantly. This confirms the prediction based on the temporal dependence of specific classifier production on noun planning: when speakers are under greater time pressure, they should produce the general classifier more often, as it can be used even when noun planning has not proceeded far enough for a specific classifier to be available.

Our findings support an availability-based production (K. Bock, 1987; V. S. Ferreira & Dell, 2000) model of speaker choice. More broadly, our results also speak to the body of work showing incremental production (W. J. Levelt, 1989; V. S. Ferreira, 1996; V. S. Ferreira & Dell, 2000; Brown-Schmidt & Tanenhaus, 2006; Brown-Schmidt & Konopka, 2015), where speakers plan as they speak and the production of an utterance is piece by piece.

6.2 Speaker-oriented, listener-oriented, or both?

From the perspective of language production, especially in real-time speech, speaker choice in classifier production reflects speaker-centric availability-based effects. Classifier choice can be seen (at least partially) as a by-product of speaker information processing.

From the perspective of language design, the redundancy of the information content shared between nouns and their classifiers (especially specific classifiers) can potentially enhance robust communication. From our corpus data (source data from (Sogou, 2008)), when multiple options are available, overall there is a preference to use a specific classifier (55.4% out of 120,861 observations in the development set and 63.1% out of 175, 988 observations in the test set used a specific classifier[1]). This general trend provides partial support for the Uniform Information Density hypothesis (Levy & Jaeger, 2007; Jaeger, 2010), given that nouns generally contain more information content than classifiers, and specific classifiers generally contains more information content than the general classifier.

Here I offer a speculation based on the existing data in this study. When possible, speakers would prefer to use a specific classifier. The use of a specific classifier offloads part of the information content from the noun, making the distribution of information content more uniformly between the classifier and the noun. However, the capacity

[1]The noun list in the development set was manually checked and filtered, but the test set was not manually checked. The higher value in the test set may be due to parsing and segmentation errors generated by the automatic NLP tools. Another thing to keep in mind is that these data are for written texts and the specific classifier ratio may be higher than that in spontaneous speech.

of using a specific classifier is constrained by production pressure, specifically, factors affecting the availability of the target specific classifiers such as noun predictability and noun frequency. If we take a step further and introduce more noise into the communication channel, what would happen? Would we maybe get some effects that reflect principles of UID in the production experiment? This remains an open question for future research.

6.3 What does classifier choice tell us about lexical access?

We asked participants to describe the quantity and the name of identical objects shown in a picture to elicit classifier production. From the classifiers elicited where counting is likely required (number of objects bigger than 5), the data suggest the following order is likely to happen during this process: count objects, retrieve noun, and select classifier. This order might be language specific and different for other classifier languages that have different word order. For example, in Japanese, the primary word order for a numeral noun phrase is Noun Numeral Classifier. It is possible and perhaps more efficient to have the following process: retrieve noun, count objects, select classifier. Currently we know very little in this area. Future work could look into this.

The effects of noun frequency on classifier choice and response latency also indicate the dependency of the classifier on the noun, suggesting that access to the noun lemma is important for producing a proper specific classifier.

It is intriguing to see that the effect of noun frequency on response latency is smaller than the effect of number, though the effect of noun frequency on specific classifier rate is much larger than the effect of number. It seems that linguistic factors carrying information about language experience such as noun frequency play an important role in speaker choice, whereas cognitive factors such as having to count (especially under time pressure) have a bigger effect on response latency and a smaller effect on speaker choice. At this moment, we do not understand this fully yet and this study opens up more new questions.

6.4 Computation vs. storage: implications from classifier use

Part of availability effects come from linguistic experience such as word frequency. We found that frequent nouns elicit more specific classifiers whereas rare nouns are more likely to to used with the general classifier. It is possible that frequently used nouns and their associated classifiers are stored as chunks. The more they are used, the stronger the connection between the classifier and the noun. For most other nouns that are not frequently used, finding their specific classifiers likely requires some computation. One of the easiest and perhaps least costly rule is to use the general classifier.

There is some evidence for storage as chunks in lexicaliztation of classifier-noun pairs. In modern Chinese (evolved through time), there exists a set of classifier-noun pairs that have been lexicalized to disyllabic nouns (Ma, 2015). In these disyllabic nouns, the first character is a noun, and the second character is a corresponding classifier (see examples (1) - (4)).

(1) 书本
shuben
book-CL.book (“book(s) in general")

(2) 一本 书
yi ben shu
one CL.book book (“one book")

(3) 车辆
cheliang
vehicle-CL.vehicle (“vehicle")

(4) 一辆 车
yi liang che
one CL.vehicle vehicle (“one car")

It is worth mentioning that the above-mentioned classifier-noun lexicalization generally only exists for very frequent nouns, and those nouns have a strong if not almost exclusive preference for a specific classifier. In this study, I focus on nouns that are compatible with both the general classifier and at least one specific classifier to express more or less the same meaning. If we think of applicable nouns as a spectrum with regard to their preference toward using a general classifier or a specific classifier, then the objects of this study lie in the middle part of this spectrum, where optionality thrives.

Looking back on the historical change in Chinese classifiers, what are implications of insights regarding online processing for theories of language change? If there are several hundred years or longer where there was variability in the word order options for classifier production (e.g., Num CL Noun versus Noun Num CL and other varieties), data from that period might shed light into this question. This is out of the scope of the current thesis, however, it would be an interesting path going forward.

6.5 Conclusion

Using evidence from naturalistic language datasets and real-time language production experiments, I have argued that speaker choice is emerged from availability-based production and is the results of speaker-centric information processing, where availability is affected by linguistic experience such as contextual predictability and word frequency, as well as real-time production pressure in speaking. From the perspective of language design, the redundancy in information content shared between nouns and classifiers provides a mechanism that potentially is beneficial for the listener, making communication more robust. However, the capacity of using a specific classifier is constrained by speaker-oriented availability production.

Appendix A

Offline Classifier Preferences

To gain a better understanding of the offline preference for classifiers, I ran an classifier knowledge survey on Witmart, the same crowdsourcing website as the one used in Experiment 1.

The Mandarin-speaking participants (N=38) were given pictures that are identical to the ones used in Experiment 1 with a brief description using a numeral and a noun, leaving the classifier blank. The participants were asked to fill in a classifier of their choice. Each item had two questions. Figures A-1 and A-2 show the sample questions for an item.

- Q1: Which classifier would you prefer to use to describe the picture?
- Q2: If you don't use the classifier in the previous question, what alternative classifier would you use? If not applicable, indicate by typing "none".

Figure A-3 shows the results from both real-time (from Experiment 1) and offline results for classifier use for the same set of items (between participants). Offline 1 refers to the first classifier came to the participant's mind (Q1). Offline 2 refers to the classifier used in offline 1 plus another alternative classifier option the participant would use if applicable (Q1 and Q2). The quick and slow conditions are conditions from Experiment 1.

Pairwise analysis reveals the following results. There was no (though marginal) significant difference in the response of classifier use between Offline 1 vs. Slow ($p =$

Figure A-1: Offline survey Question 1: Which classifier would you prefer to describe the picture above?

Figure A-2: Offline survey Question 2: If you don't use the classifier you just chose, what alternative classifier would you use? If not applicable, indicate by typing "none".

0.051). However, the difference between Offline 2 vs. Slow is significant ($p < 0.001$), so does Offline 2 vs. Offline 1 ($p < 0.001$). This indicates that some part of classifier knowledge or preferences may not be seen during real-time production. However, speakers do have access to the knowledge of specific classifier offline.

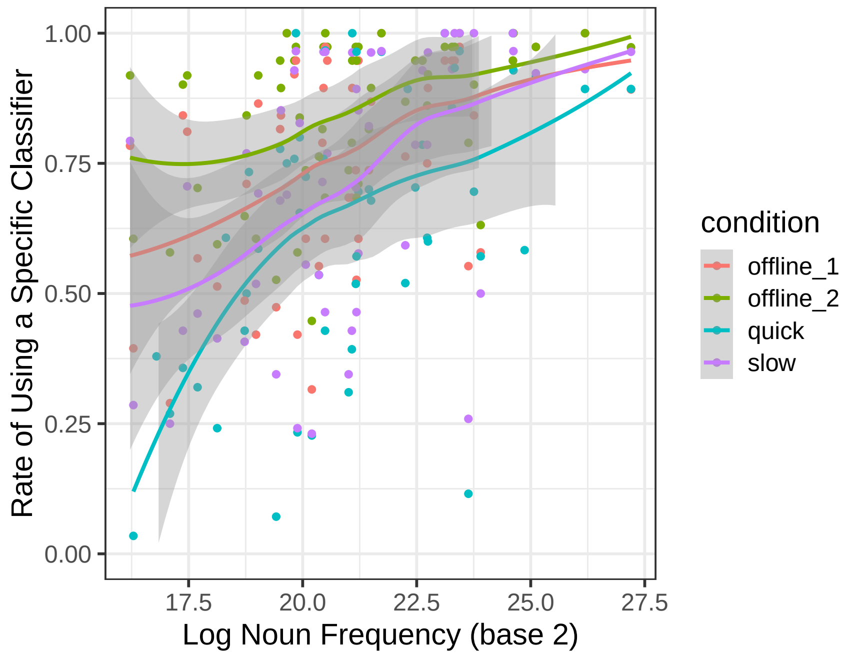

Figure A-3: Offline and real-time results: the relationship between noun frequency and rate of using a specific classifier.

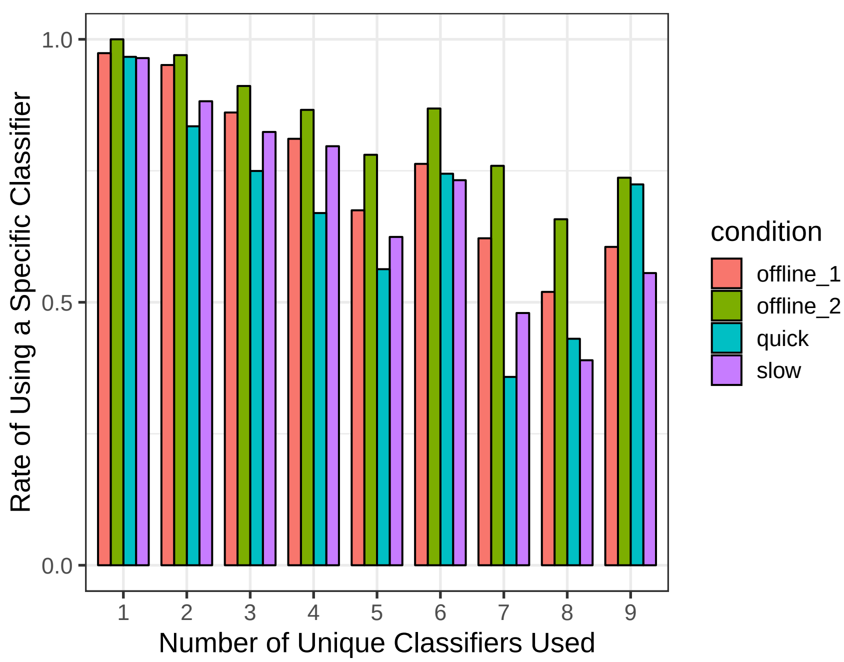

Figure A-4: Offline and real-time results: number of unique classifiers used and rate of using a specific classifier

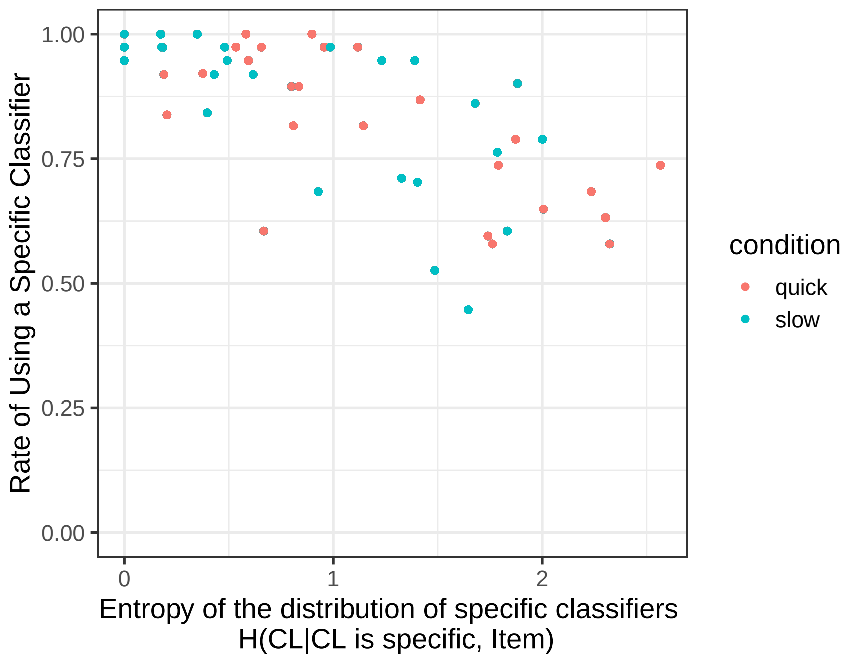

Figure A-5: Rate of using a specific classifier in real-time production and classifier entropy in offline data

Would it be possible that it is harder for speakers to choose a specific classifiers for nouns associated with more specific classifiers? I calculated the number of unique classifiers used data and the entropy of the distribution of specific classifiers using the offline 2 data. Figures A-4 shows the number of unique classifier used against the rate of using a specific classifier. Figure A-5 shows the entropy of the distribution of specific classifiers (offline) against the rate of using a specific classifier (real-time).

If we take the offline 2 specific classifier rate as a baseline, above and beyond this baseline, is the entropy of the distribution of specific classifiers (based on the offline 2 data) $H(\text{CL}|\text{CL is specific}, \text{Item})$ predictive of the real-time specific classifier usage? Translating this question to a logistic regression, the full formula in the style of `R`'s `lme4` is:

```
classifier_choice ~ offline2_specific_classifier_rate + entropy + condition
+ (1 + offline2_specific_classifier_rate + entropy| participant)
+ (1 + condition| item )
```

As shown in Figure A-5, numerically, the direction of entropy effect on classifier use is: if an item has a higher entropy of specific classifier distribution, then it is less likely a specific classifier would be used with that item in real-time production. The numerical trend is aligned with the idea that competition among different specific classifiers could potentially impact the availability of specific classifiers in real-time language production in addition to other effects investigated in the main body of this thesis. However, the results of the above analysis did not find a main effect of specific classifier entropy ($p = 0.240$) above and beyond the offline specific classifier baseline ($p < 0.001$) and the main effect of condition ($p < 0.05$). For now, the data shows that classifier entropy does not have an effect on classifier production above and beyond the speaker's offline classifier preference.

References

Aikhenvald, A. Y. (2000). Classifiers: A typology of noun classification devices. *Studies in Typology and Linguistic Theory. Oxford: Oxford University Press*.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Allan, K. (1977). Classifiers. *Language*, 285–311.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, *62*(1), 67–82.

Aronoff, M. (1976). *Word formation in generative grammar (linguistic inquiry monograph 1)*.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. , *47*(1), 31–56.

Baayen, R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, *9*(3), 447–469.

Baayen, R. H. (2002). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

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

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production. *Journal of Memory and Language*, *24*(1), 89–106.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive psychology*, *17*(3), 364–390.

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

Barth, D. (2019). Effects of average and specific context probability on reduction of function words be and have. *Linguistics Vanguard*, *5*(1).

Behaghel, O. (1930). Von deutscher wortstellung. *Zeitschrift für Deutschkunde*, *44*(1930), 81–89.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, *18*(3), 355–387.

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, *26*(2), 119–137.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016a). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, *157*, 81–94.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016b). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, *157*, 81–94.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, *5*(4), 325–337.

Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic bulletin & review*, *22*(3), 833–843.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Buck, C., Heafield, K., & Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *Lrec* (Vol. 2, p. 4).

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change*, *14*(3), 261–290.

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). John Benjamins Publishing.

Cheng, L. L.-S., & Sybesma, R. (1998). Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, *28*(3), 385–412.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1–33.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.

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

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive psychology*, *37*(3), 201–242.

Cole, R. A., & Perfetti, C. A. (1980). Listening for mispronunciations in a children's story: The use of context by children and adults. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 297–315.

Da, Z. (2004). *Shanggu hanyu shuliangci yejiu.*

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes. *Phonetics and phonology in language comprehension and production: Differences and similarities*, *6*, 9–37.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Dixon, R. M. (1986). Noun classes and noun classification in typological perspective. *Noun classes and categorization*, 105–112.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/`

Ferreira, F., & Swets, B. (2002). How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84.

Ferreira, V. S. (1996). Is it better to give than to donate? syntactic flexibility in language production. *Journal of memory and language*, *35*(5), 724–755.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340.

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of verbal learning and verbal behavior*, *18*(1), 1–20.

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology*, *33*(4), 465–495.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, *22*(6), 829–834.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, *44*(3).

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748–775.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and* *language*, *66*(4), 789–806.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51.

Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of chinese individual classifiers. *Language and Cognitive Processes*, *24*(7-8), 1124–1179.

Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the london-lund corpus of spontaneous conversation. *Linguistics*, *19*(7-8), 805–818.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In (pp. 199–206).

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*.

Gil, D. (2013). Numeral classifiers. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/55`

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, *22*(1), 131–153.

Goldberg, A. E. (2016). Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption. *Language and cognition*, *8*(3), 369–390.

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

Greenberg, J. H. (1972). Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. *Working Papers on Language Universals*, *9*.

Gundel, J. K. (1988). Universals of topic-comment structure. *Studies in syntactic typology*, *17*(1), 209–239.

Habibi, A. A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies* (pp. 1–8).

Hawkins, J. A. (1994). *A performance theory of order and constituency* (Vol. 73). Cambridge University Press.

Hawkins, J. A. (2014). *Word order universals* (Vol. 3). Elsevier.

He, J. (2003). *Liang ci yi dian tong [the prompt understanding of measure words]*. Beijing Language and Culture University Press.

Hiranburana, S.-a. (1979). A classification of thai classifiers. *Southeast Asian Linguistic Studies*, *4*, 39–54.

Huang, Z. (1964). Cong jiawen, jinwen de yingyong, kaocha hanyu liangci de qiyuan yu fazhan. *Zhongguo Yuwen*.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, *20*(2), 137–194.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language*, *45*, 229–254.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498–525.

Kiparsky, P. (1993). Blocking in nonderived environments. In *Studies in lexical phonology* (pp. 277–313). Elsevier.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Klein, N. M., Carlson, G. N., Li, R., Jaeger, T. F., & Tanenhaus, M. K. (2012). Classifying and massifying incrementally in chinese language comprehension. *Count and mass across languages*, 261–282.

Kovalenko, L. Y., Chaumon, M., & Busch, N. A. (2012). A pool of pairs of related objects (poporo) for investigating visual semantic integration: behavioral and electrophysiological validation. *Brain Topography*, *25*(3), 272–284.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, *83*, 152–178.

Levelt, W., & Maasen, B. (1981). Lexical search and order of mention in sentence production. In W. Klein & W. J. M. Levelt (Eds.), *Crossing the boundaries in linguistics* (pp. 221–252). Springer.

Levelt, W. J. (1989). *Speaking: From intention to articulation.* MIT press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2018). Communicative efficiency, uniform information density, and the rational speech act theory. In *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 684–689).

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Li, P., Barner, D., & Huang, B. H. (2008). Classifiers as count syntax: Individuation and measurement in the acquisition of mandarin chinese. *Language Learning and Development*, *4*(4), 249–290.

Li, P., Huang, B., & Hsiao, Y. (2010). Learning that classifiers count: Mandarin-speaking children?s acquisition of sortal and mensural classifiers. *Journal of East Asian Linguistics*, *19*(3), 207–230.

Li, W. (1980). Hanyu shigao.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, *6*(3), 172–187.

Liu, S. (1965). *Weijin nanbei chao liangci yanjiu [a study on classifiers in wei-jin nan-bei dynasty].* Zhonghua Shuju.

Lohmann, A. (2011). Help vs help to: a multifactorial, mixed-effects account of infinitive marker omission. *English Language & Linguistics*, *15*(3), 499–521.

Ma, A. (2015). *Hanyu geti liangci de chansheng yu fazhan[the emergence and development of chinese individual classifiers].* China Social Sciences Press.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, 226.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, *15*(1), 19–44.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Acl (system demonstrations)* (pp. 55–60). Retrieved from http://www.aclweb.org/anthology/P14-5010

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, i–178.

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes*, *3*(1), 1–16.

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In *Advances in neural information processing systems* (pp. 1585–1593).

McClelland, J. L., & O'Regan, J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 634.

Meng, F. (2009). Liangci tiao de chansheng ji qi lishi yanbian. *Journal of Ningxia University (Humanities & Social Sciences Edition*, *31*(1), 35–40.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological bulletin*, *134*(3), 427.

Pinker, S. (1999/2011). *Words and rules: The ingredients of language*. Harper Prennial.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43–102.

Prince, E. F. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–256). New York: Academic Press.

Rayner, K., Binder, K. S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, *41*(7), 943–954.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sogou. (2006). *Sogou lab data: Internet lexicon 2006 version.* `https://www.sogou.com/labs/resource/w.php`. (Accessed: 2018-11-30)

Sogou. (2008). *Sogou lab data: Sohu news corpus 2008 version.* `http://www.sogou.com/labs/resource/cs.php`. (Accessed: 2017-05-30)

Srinivasan, M. (2010). Do classifiers predict differences in cognitive processing? a study of nominal classification in mandarin chinese. *Language and Cognition*, *2*(02), 177–190.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*(1), 1.

Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: Evidence from either... or. *Journal of experimental psychology: Learning, memory, and cognition*, *32*(2), 425.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Interspeech* (Vol. 2002, p. 2002).

Sudo, Y. (2018, Decenber). *Lecture notes in plin0020 advanced semantic theory: Classifier languages.* University College London.

Tai, J., & Wang, L. (1990). A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association*, *25*(1), 35–56.

Tai, J. H. (1992). Variation in classifier systems across chinese dialects: towards a cognition-based semantic approach. *Chinese Language and Linguistics*, *1*, 587–608.

Tai, J. H. (1994). Chinese classifier systems and human categorization. *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change*, 479–494.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Temperley, D. (2007). Minimization of dependency length in written english. *Cognition*, *105*(2), 300–333.

Traxler, M. J., & Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1266.

Warren, P. (2013). *Introducing psycholinguistics.* Cambridge University Press.

Wasow, T., Jaeger, T. F., & Orr, D. M. (2011). Lexical variation in relativizer frequency. In H. J. Simon & H. Wiese (Eds.), *Expecting the unexpected: Exceptions in grammar* (pp. 175–96). De Gruyter Mounton.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and cognitive processes*, *19*(6), 713–755.

Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic tac toe: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548–1557.

Xiandai, H. C. R. G. (2008). *Xiandai hanyu changyongcibiao.[lexicon of common words in contemporary chinese].* The Commercial Press.

Xun, E., Rao, G., & Zang, J. (2016). The construction of the bcc corpus in the age of big data. *Corpus Linguistics*, *3*(1), 93–118.

Yang, C. (2005). On productivity. *Linguistic variation yearbook*, *5*(1), 265–302.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language.* MIT press.

Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 17th annual conference of the north american chapter of the association for computational linguistics: Human Language Technologies* (pp. 1997–2005).

Zhan, M., & Levy, R. (2019). Availability-based production predicts speakers' real-time choices of Mandarin classifiers. In *Proceedings of the 41st annual meeting of the cognitive science society* (p. 1268-1274).

Zhang, H. (2007). Numeral classifiers in mandarin chinese. *Journal of East Asian Linguistics*, *16*(1), 43–59.

Zhu, M., Zhang, Y., Chen, W., Zhang, M., & Zhu, J. (2013). Fast and accurate shift-reduce constituent parsing. In *Acl (1)* (pp. 434–443). Retrieved from `http://www.aclweb.org/anthology/P13-1043`

Zipf, G. K. (1949). Human behavior and the principle of least effort.