Lexical processing in simplified Chinese: an investigation using a new large-scale lexical database

D i s s e r t a t i o n
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Ching Chu Sun

aus

Taipei, Taiwan

2016

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Harald Baayen

Mitberichterstatter: Prof. Dr. Achim Mittag

Mitberichterstatterin: Prof. Dr. Laurie Feldman

Tag der mündlichen Prüfung: den 7. November 2016

To my daughters 忻之 and 希之,

To my cats Nini and Mimi,

And to my husband Peter,

Acknowledgements

I would like to thank my supervisor, Harald Baayen, who is the most positive and encouraging person I have met in my life. Harald was always patient and gave me answers to the countless questions I had. I often wondered how he managed to be that positive and never get upset about anything. Because of Harald, I never gave up on learning new techniques. Harald once told me that as long as one is willing to work hard and be persistent, intelligence and background do not play a decisive role in one's academic career. I remembered what he said and did my best to succeed in my research goals, because the last person I wanted to disappoint was Harald.

I would also like to thank Harald for finding Dr. Achim Mittag to be my second supervisor. Dr. Mittag provided me with an opportunity to look at my data from a different perspective. His advice and suggestions about how the Chinese Lexical Database can be used for teaching Chinese were very useful and have given me many new ideas about what I can do to build a bridge between psycholinguistic research and teaching Chinese.

Harald would like to keep the tradition that his PhD students always have someone from North America on their committee. As a result, I was lucky enough to have Dr. Laurie Feldman as a third reviewer of my dissertation. After reading my thesis, Laurie pointed out how useful the Chinese Lexical Database is for designing stimuli for psycholinguistic research. She also pointed out that the work presented in my dissertation can help improve reading instruction in China. While I was working on my thesis, I did not see these opportunities myself. Therefore, I would like to thank Laurie for seeing value in my dissertation that I was unable to see myself.

I furthermore thank Dr. Andrea Weber and Dr. Britta Stolterfoht for being on my committee, and for their interesting questions and suggestions during the defense. I also thank Tineke, our group secretary. She has been very helpful and

finished many administrative procedures within an incredibly short period of time. Tineke was always accessible when I had questions about documents or about using the lab. I don't think that anyone could ever be as friendly as Tineke as a secretary and a lab manager.

Next, I would like to thank Jianqiang. In addition to being the participant in my word naming experiment, Jianqiang helped me calculate a number of measures for the Chinese Lexical Database. I also thank Lea, who helped me run many participants. She was a pleasure to work with. In addition, a special thank you goes to my dear friends Robert, Martijn, Cyrus, and Bärbel for their support and their confidence in me over the years.

The next words of gratitude might sound a bit funny. However, I would like to thank my cats Nini and Mimi as well. My beloved cat Nini came with me from Taiwan to Canada and from Canada to Germany. I always felt happy and at peace when Nini was around. He was with me for twenty years. He was there when I finished my MA degree in Taiwan; he was also there when I finished my MSc degree in Canada. Unfortunately, he was not around when I finished this dissertation. I am always grateful for the joy he brought me. Mimi is a sweet sweet cat. I want to thank her for the happiness she has brought to my family.

In addition, I would like to thank my children Delight and Hope. Because of me, they had to learn four different languages in nine years. Delight lost her friends in Taiwan and Canada. In the beginning, she did not speak a word of German and life in Germany was not easy for her. She has, however, adapted very well. Furthermore, she has been helping around in the house and has given me a lot of support. Similarly, Hope did not speak a word of German either when we just arrived in Germany. Life was also not easy for her. However, in addition to doing well in school, Hope has been a big help in the house. She can do dishes, laundry, and even go shopping. Because of Hope's help, I was able to spend more time on writing the thesis and preparing for the oral defense. Now, it is time to start planning some family trips.

Finally, I thank my dear husband Peter. I learned many news things by working together with Peter. He is a good teacher in statistics and coding. He would always answer my questions, even when these questions were asked over and over. He was capable of answering the same questions from different angles and therefore giving me clear explanations. Peter, the life with you in Germany has been very fruitful and enjoyable. I love you.

Contents

List of Tables

List of Figures

1 Introduction

This dissertation investigates lexical processing in Mandarin Chinese through a number of psycholinguistic experiments, using a new, large-scale lexical resource: the Chinese Lexical Database (CLD). Prior to presenting this lexical database (Chapter 2) and leveraging it to gain insight into the behavioural correlates of lexical processing in word naming (Chapter 3) and sentence reading (Chapter 4) experiments, however, I give a synoptic overview of the main topics that are under investigation in this thesis. Below, I first provide a brief introduction to Mandarin Chinese and a number of key characteristics of the language that are relevant to this dissertation. Next, I discuss some of the main findings in the psycholinguistic literature for Chinese. Finally, I outline the contents of the remaining chapters of this thesis.

1.1 A brief introduction to Mandarin Chinese

Mandarin Chinese, which is also referred to as 普通话 ("common language"), is part of the Sino-Tibetan language family. It is the official language of China and has, according to recent estimates, nearly a billion (935 million) native speakers – or 14.1% of the world population (Parkvall, 2007). By comparison, the most studied language in the field of psycholinguistics, English, had about 365 million native speakers, which is 5.52% of the world population.

Chinese is a tonal language. The basic phonological unit is the syllable. Each syllable consists of vowels and consonants in a (C)V(C) structure at the segmental level and a tone at the suprasegmental level (C. Sun, 2006). The tonal system comprises 4 basic tones and 1 neutral tone. In the writing system, syllables correspond

to 汉字 (Hanzi, literal translation: "Chinese characters"). According to the Table of General Standard Chinese Characters, there are about 8,100 Chinese characters, of which 6,500 are commonly used (Ministry of Education of the People's Republic of China, 2013).

The origin of Chinese characters dates back to about 1200 B.C. and is founded on oracle bones from that period. The oracle bone scripts developed into traditional Chinese characters. Due to the complicated nature of traditional Chinese characters, learning to read or write in Chinese is a difficult task. To master the language, one has to memorize and constantly practice thousands of characters. To improve literacy, the Chinese government decided to simplify over 2,200 characters (see Honorof & Feldman, 2006) in the 1950s. This resulted in the writing system studied in this dissertation, which is commonly referred to as simplified Chinese.

A few patterns can be observed in the simplification of Chinese characters. For some characters the original shape of the characters was retained, while the number of strokes was reduced. For instance, 愛 ("love") was simplified to 爱, and the simplified form of 鳳 ("phoenix") is 凤. The amount of simplification varies between characters. At times, this type of simplification was taken to an extreme and only one of the visual components of a character was retained. For instance, the character 産 ("to produce") was simplified to 产, and 廠 ("factory") to 厂. For some characters that were composed of two visual components in traditional Chinese, one of the components was simplified, whereas the other was not. For the character 僅 ("only", components: 亻 and 堇), for instance, the component 亻 remained the same, whereas 堇 was simplified to 又. The simplified version of the character 僅 thus is 仅.

The Chinese writing system uses a series of different strokes to build up characters. In 1965 and 1988, the Chinese government published the Modern Chinese General Character List (现代汉语通用字表; Ministry of Education of the People's Republic of China, 1988) and stipulated five basic types of strokes: 横 (一) horizontal, 竖 (丨) vertical, 撇 (丿) left-falling, 点 (丶) dot, and 折 (乚) turning strokes (National Language Commission of China (国家语委), 1997). When writing a character, the strokes are written in a certain order. The basic principle is to start a stroke from left to right, from top to bottom, or from outside to inside. The number of strokes in a character and the order of the strokes are important information for looking up words in a dictionary.

Strokes combine into visual components. Typically, the term component refers to a unit that contains semantic or phonological information, although this is not necessarily the case. Some visual components can be independent characters, whereas others cannot. Chinese characters contain one or more components. For instance, characters like 一 ("one"), 山 ("mountain"), or 口 ("mouth") consist of a single visual component. The characters 林 (木 plus 木; "trees"), 好 (女 plus 子; "good") and 家 (宀 plus 豕; "home") comprise two components. Examples of characters that consist of three components are 碧 (王 plus 白 plus 石; "jade"), 冠 (冖 plus 元 plus 存; "crest"), and 狱 (犭 plus 讠 plus 犬; "jail").

The same visual component can occur in different positions in different characters. Consider, for example, the character 山 ("mountain"), which is an independent character as well as a component in other characters. The character 山 occurs in four different positions within a character 峰 (left; 山 plus 夆; "apex"), 仙 (right; 亻 plus 山, "immortal"), 岁 (top; 山 plus 夕; "years old"), and 岔 (bottom; 分 plus 山; "to diverge"). Not all components can appear in different positions, however. The component 艹 ("grass"), for instance, occurs at the top of a character only (e.g., in the characters 草 (艹 plus 早; "grass") and 莊 (艹 plus 壯; "strong")). Another example of a component that appears in a single position only is 刂 ("knife"). This component is always located at the right side of a character (e.g., in the characters 到 (至 plus 刂; "arrive") and 刊 (干 plus 刂; "to publish")).

One type of visual component is used as an index to look up words in a dictionary. This type of component is called 部首 ("section-head"), which is typically translated as "semantic radical". Most semantic radicals are independent characters as well. Characters in which a semantic radical occurs typically are semantically related to the radical. For instance, the character 口 ("mouth") is used as a semantic radical in 吃 ("eat") and 喝 ("drink"). In this case, there is a direct relation between the meaning of the semantic radical ("mouth") and the meanings of the characters it occurs in ("eat" and "drink"). Similarly, the character 贝 ("sea shell") is used as a semantic radical in characters like 财 (贝 plus 才; "fortune") and 贪 (今 plus 贝; "greedy"). The meaning of the semantic radical 贝 seems unrelated to the meanings of the characters 财 and 贪 from a contemporary perspective. However, there is a historic semantic connection between sea shells and monetary value: sea shells were used as a currency in ancient Chinese society.

Although the meaning of characters is typically related to the meaning of the semantic radical, this is not always the case. For instance, the radical 冫 ("ice") occurs in the characters 冰 and 准. The meaning of the character 冰 is identical to that of the semantic radical: "ice". However, the meaning of the character 准 is "to allow". Therefore, the character 准 is semantically unrelated to the semantic radical 冫. While the semantic radicals 贝 ("sea shell") and 口 ("mouth") can be used as independent characters, the semantic radical 冫 ("ice") is an example of a semantic radical that cannot occur as an independent character.

A second type of visual component provides information about the pronunciation of a character. This type of component is called 声旁 (literal meaning: "sound side") and is commonly referred to as "phonetic radical". The pronunciation of a character in which a phonetic radical occurs can be identical to the pronunciation of the phonetic radical. The character 铜 ("copper"; "[tʰɔŋ2]"), for instance, has the same pronunciation as its phonetic radical 同: "[tʰɔŋ2]". This, however, is not necessarily the case. The pronunciation of the character 岗 ("a small hill"; "[gɑŋ3]"), for instance, differs from the pronunciation of its phonetic radical 冈 ("[gɑŋ1]") at the suprasegmental level. The pronunciation of the character 姝 ("[ʂu1]"), by contrast, differs from the pronunciation of its phonetic radical 朱 ("[tʂu1]") at the segmental level. Pronunciations of the character and the phonetic radical can differ at both the segmental and the suprasegmental level as well. The phonetic radical 工 of the character 红 ("red";"[hɔŋ2]"), for instance, is pronounced as "[gɔŋ1]".

About a third of the words in Mandarin Chinese are single-character words (Honorof & Feldman, 2006). Examples of one character words are 米 ("rice") and 走 ("to walk"). Although words that consist of more than two character exist, an overwhelming majority of the remaining two thirds of Chinese words consist is a combination of two characters. Characters can be combined in different ways to form two-character words. In one type of two-character words, characters with similar meanings are combined to form a two-character word. Examples of this type of word formation, in decreasing order of semantic similarity of the component characters, are 朋友 ("friend-friend"; "friend"), 价值 ("value-price"; "value"), and 天地 ("sky-ground"; "world"). A special case of this type of word formation is reduplication. Repeating the character 人 ("person"), for instance, yields the two-character word 人人 ("everyone").

In another type of two-character words, the first character semantically modifies the second character. Combining the characters 蛇 ("snake") and 行 ("walk"), for instance, results in the two-character word 蛇行, which means "to zigzag". Similarly, the first character 渺 ("tiny") modifies the meaning of the second character 视 ("to look at") in the two-character word 渺视 ("to despise"). Furthermore, two semantically unrelated characters can combine into a two-character word to generate a new meaning. The characters 吃 ("to eat") and 醋 ("vinegar"), for instance, combine into the word 吃醋, which means "jealous". Combining the characters 民 ("people") and 主 ("to master") yields the meaning "democracy" for the two-character word 民主.

This concludes my brief introduction to Mandarin Chinese. I discussed a number of key properties of the language that are pivotal for an overall appreciation of the work presented in this dissertation. There are many more interesting characteristics of Chinese that were not explicitly discussed here. Some of the characteristics are pertinent to specific parts of this thesis. Whenever this is the case, the relevant property of the language is discussed in the corresponding section of this dissertation.

1.2 Psycholinguistic research in Chinese

The field of psycholinguistics investigates the psychological processes that underlie the acquisition, the comprehension and the production of languages, both through behavioural experiments and through computational simulations of language processing. Traditionally, psycholinguistic research focused primarily on Indo-European languages. Psycholinguistic studies of Mandarin Chinese were few and far between until the 1980s. Over the last decades, however, the study of non-alphabetical languages in general and of Mandarin Chinese in particular has gained increased popularity. This resulted in a large number of experimental studies investigating different aspects of language processing in Chinese. Below, I discuss a number of key findings in psycholinguistic studies of Mandarin Chinese. For more information, I refer the interested reader to three books that provide a more comprehensive overview of psycholinguistics in Chinese: Language Processing in Chinese (H. C. Chen & Tzeng, 1992), Reading Chinese Script: a Cognitive Analysis (J. Wang et al., 1999), and The Handbook of East Asian Psycholinguistics (Li et al., 2006).

The most basic psycholinguistic finding across languages is the word frequency effect: words that occur often in the language are processed faster than rare words. The word frequency effect is solidly established for Mandarin Chinese. Both one-character words (Seidenberg, 1985; Y. Liu et al., 2007) and two-character words (I. M. Liu, 1999) are named (i.e., read aloud) faster when they are more frequent. Similarly, reaction times in lexical decision experiments – in which participants are asked to indicate whether or not a character or word exists in Chinese – are shorter for high frequency one-character (Lee et al., 2015; Sze et al., 2014) and two-character words (Zhang & Peng, 1992; Peng et al., 1999).

Not only the frequency of words influences behavioural measures of lexical processing, but also the frequency of the characters within words. Zhang and Peng (1992), Taft et al. (1994), and Peng et al. (1999) all found character frequency effects in lexical decision. Character frequency effects were observed in other measures of language processing as well: G. Yan et al. (2006) observed a character frequency effect on eye fixation durations on two-character words and Kuo et al. (2003) and Lee et al. (2004) found character frequency effects in an fMRI study.

The visual complexity of characters likewise influences lexical processing. The effect of the visual complexity is present at different grain sizes. At a small grain size, characters with a high number of strokes yield longer reaction times in lexical decision (Lee et al., 2015) and longer naming latencies (Y. Liu et al., 2007; Leong et al., 1987) as compared to characters that consist of fewer strokes. At a larger grain size, a greater number of visual components leads to longer naming latencies (Y. Liu et al., 2007).

A third type of effect documented in the literature is related to a lexical variable that is typically referred to as family size in English (see e.g., Baayen et al., 2006; Schreuder & Baayen, 1997). In Mandarin Chinese, I define family size as the number of words a character occurs in. Y. Liu et al. (2007) referred to family size as "number of word formations" and found that characters that occur in two character words are named faster. Tsai et al. (2006) used the term "neighbourhood size" to describe position-specific family size (i.e., the number of words in which a character occurs in a specific position) and found shorter lexical decision latencies for two-character words with an initial character that occurred in many other two-character words. The frequency of a word's family members, the family frequency, influences lexical processing as well. In a lexical decision ERP experiment Huang et al. (2006) found inhibitory effects of family frequency, with longer naming latencies and a greater N400 for words that contained characters with high family frequencies.

As noted above, Chinese has an inventory of about 8,100 characters (Ministry of Education of the People's Republic of China, 2013). These characters are mapped onto a limited set of phonological forms. According to estimations by DeFrancis (1984) (as cited by J. Y. Chen & Dell, 2006), there are about 1,200 unique syllables when tone is taken into consideration. When tone is ignored, this number is reduced to about 400. A large number of orthographic units thus is mapped onto a relatively small number of phonological forms. As a result, the mapping between orthography and phonology is less than consistent.

Homophony occurs when two characters have the same pronunciation, but are orthographically distinct. Despite the fact that many effects of homophony have been observed, the qualitative nature of the influence of homophony on lexical processing remains unclear. Lee et al. (2015) and W. Wang et al. (2012) found inhibitory effects in visual and auditory lexical decision, respectively. By contrast, a number of other studies found facilitatory effects in word naming (Ziegler et al., 2000) and in auditory word recognition (H. C. Chen et al., 2009; W. F. Chen et al., 2016). Finally, Y. Liu et al. (2007) did not find an effect of the number of homophones in a multiple regression study that controlled for the effects of a large number of other lexical variables.

At the word level, studies of the consistency between orthography and phonology have focused primarily on the consistency of the phonology-to-orthography mapping (i.e., homophony). Below the character level, however, the consistency of the orthography-to-phonology mapping has been shown to influence lexical processing. Characters with pronunciations that are identical to the pronunciation of their phonetic radicals are named faster than characters for which the pronunciation of the character differs from the pronunciation of the phonetic radical (Y. Liu et al., 2007; Seidenberg, 1985; Hue, 1992).

Regularity is a binary variable: the pronunciation of a character is either identical to the pronunciation of the phonetic radical or not. Glushko (1979) proposed a more sophisticated, numerical measure of the reliability of the orthography-to-phonology mapping: consistency. Given the set W of all words that contain a phonetic radical p, the consistency of the phonetic radical p is defined as the proportion of words in W with pronunciations identical to the pronunciation of p. Fang et al. (1986) demonstrated that naming latencies for one-character words are faster for characters that contain more consistent phonetic radicals. An effect of consistency was also observed in ERP (Hsu et al., 2009) and fMRI (Lee et al., 2004) studies.

In addition to the reliability of the information provided by the phonetic radical, recent studies have documented effects of the reliability of the information provided by the semantic radical as well. M. J. Chen and Weekes (2004) and M. J. Chen et al. (2006) found that characters with meanings that are the same as or similar to the meaning of the semantic radical are processed more efficiently than characters with meanings that are dissimilar to the meaning of the semantic radical are processed more efficiently in semantic categorization and lexical decision tasks.

A second type of effect below the character level is the effect of family size. The more characters a phonetic radical occurs in (i.e., the greater its family size), the faster the reaction times in lexical decision experiments (Feldman & Siok, 1997; Taft & Zhu, 1997; Lee et al., 2015). Recently, the family size of the phonetic radical has also been shown to influence event-related potentials in a reading task (Hsu et al., 2009). The family size effect at the radical level is not limited to phonetic radicals. In a series of studies, Feldman and Siok demonstrated that a similar facilitatory effect of radical family size in lexical decision is present for semantic radicals as well (see e.g., Feldman & Siok, 1997, 1999a, 1999b).

Effects of lexical properties of the phonetic and semantic radical have been interpreted as evidence for compositional processing at the character level. Taft (2006), for instance, proposed a reading model of Chinese in which access to visual components precedes access to characters (see also Taft et al., 1999; Taft & Zhu, 1997). As noted by Feldman and Siok (1999a), this view stands in contrast to theories that assume that the character is the "primary unit of visual recognition (e.g. Cheng, 1981; Hoosain, 1991; I. M. Liu, 1988)". The visual components at the radical level themselves, Taft (2006) suggested, are activated by visual information at the stroke level.

Taft (2006) furthermore argued that the orthography to phonology mapping is mediated by lexico-semantic representations. Taft (2006) therefore proposed a single route between orthography and phonology. By contrast, based on the observation that – at least under some experimental conditions – phonological priming effects precede semantic priming effects, Perfetti and Tan (1998), Perfetti and Tan (1999), and Perfetti and Liu (2006) argued that a second, direct route from orthography to phonology exists as well. The investigation of lexical processing in Chinese thus provides an additional perspective on the single versus dual route debate in the psycholinguistic literature in English (see e.g., Coltheart et al., 2001; Harm & Seidenberg, 2004).

1.3 Outline of this dissertation

Above, I presented some of the key findings for Chinese in the psycholinguistic literature. These findings were framed in terms of the influence of lexical distributional variables on behavioural measures of language processing and were related to orthographic, phonological and semantic properties of words, characters and radicals. The influence of these variables was typically established in isolated experiments targeted at the effect of a specific lexical predictor and – in the absence of a large-scale lexical database for Chinese – variables were nearly always calculated independently for experiments carried out by different researchers. This practice has at least two potential ramifications for the reliability, comparability and replicability of experimental results.

First, the independent calculation of lexical predictors for individual experiments leads to substantial differences between the information captured by the same measures in different experiments. While Cai and Brysbaert (2010) found correlations of $r \approx 0.80$ between word frequencies measures based on movie subtitles and frequency measures based on existing corpora of Chinese, for instance, I found reduced correlations of $r \approx 0.65$ between the subtitle frequencies of Cai and Brysbaert (2010) and word frequency measures obtained from two large-scale corpora of written Chinese (see Chapter 2).

The differences in the information captured by conceptually identical measures derived from different resources are an unwanted source of variance in behavioural measures of language processing. As noted above, for instance, the qualitative nature of the homophony effect differs between experiments. To a large extent, this is likely to be due to differences in the experimental task and the experimental design. To some extent, however, discrepancies between the results of studies may also be a consequence of the inconsistent calculation of the same lexical predictor. The availability of unified, well-documented lexical predictors that are publicly available could help overcome this problem.

Second, due to the unavailability of a large-scale lexical resource, it is difficult and time-consuming to control for the effects of a large set of lexical predictors that are extrinsic to the experimental question, but that may nonetheless have an influence on the dependent variable. Researchers tend to control for a limited set of lexical variables that are known to influence the experimental task at hand and that are relatively easy to calculate, such as word and character frequency and stroke

count. Predictors for which the effects are less well-established, or that are harder to retrieve from a corpus are – understandably – often disregarded. A large-scale lexical database from which lexical predictors can readily be extracted can help establish the influence of a large set of predictor in an efficient and straightforward manner, for instance in a multiple regression design.

The largest lexical resource for simplified Chinese that is currently available is a database compiled by Y. Liu et al. (2007). This database is a valuable resource that contains naming latencies, as well as 15 lexical predictors for $2,423$ one-character words. In Chapter 2 of this dissertation, I present a lexical database that is an order of magnitude larger than existing resources for Mandarin Chinese. This lexical database, the Chinese Lexical Database (henceforth CLD), contains 141 numerical variables and 23 categorical variables for $30,645$ words ($4,710$ one-character words and $25,935$ two-character words).

The lexical predictors in the CLD describe lexical properties of Chinese at the word level, the character level and the sub-character level. The database includes wide range of measures related to the frequency, the visual complexity, the phonology, and the mapping between the orthography and the phonology for a word. Furthermore, the CLD contains a number of information-theoretic measures. Thus far, the role of the information-theoretic structure of the language has received relatively little attention in psycholinguistic studies of Mandarin Chinese. In the context of the discrimination learning framework, however, Baayen et al. (2011) have demonstrated that information-theoretic properties of the language have a considerable influence in language processing. Therefore, it is interesting to establish to what extent these measures help better understand lexical processing in Chinese as well.

Chapter 3 is a first exploration of the predictive power of the lexical distributional variables in the CLD. I present the results of a word naming experiment, in which we asked a native reader of simplified Chinese to name all $30,645$ words in the CLD. During the experiment, naming latencies, as well as pronunciation durations and the durations of fixations of the eye were recorded. I investigated the quantitative contribution of the predictors in the CLD using a machine learning technique and established the qualitative nature of the effects of these predictors using non-linear regression models. The results of this word naming experiment provide interesting new insights into lexical processing in the word naming task in Mandarin Chinese and include hitherto unobserved effects of a number of information-theoretic measures.

While Chapter 3 focuses on language processing at and *below* the word level, Chapter 4 gauges the lexical properties that influence processing at and *above* the word level through phrase reading and sentence reading experiments. More specifically, I investigated how locative phrases – the Chinese equivalent of prepositional phrases (i.e., "on the table") in English – are processed. In a translation verification task, participants read locative phrases in isolation or in sentence contexts. During the experiment eye movement patterns were recorded. I analyzed these eye movement patterns using non-linear regression models and demonstrated that a dynamic search for information allows for highly efficient sentence reading in Chinese. I found phrase-level effects that have not been documented for Chinese before, again including effects of a number of information-theoretic measures.

I conclude this dissertation with a brief overview of the contributions of this thesis of the psycholinguistic literature for Chinese and a discussion of outstanding issues and promising areas we might explore in future research. I argue that the Chinese Lexical Database (CLD) is a valuable resource for psycholinguistic research in Chinese that can easily be extended on the basis of future experimental research or requests from interested colleagues. I furthermore argue that information-theoretic approaches offer a promising framework for a more comprehensive understanding of lexical processing in Chinese, both at and above the word level. I use the "I"-form in the Introduction and in the Conclusions section of this dissertation. However, as I stand on the shoulders of giants I will use the "we"-form in the main body of the text.

2

Chinese lexical database

2.1 Introduction

Over the last decades, the wealth of experimental research in the psycholinguistic literature has been complemented with studies that have made available large-scale lexical resources for a number of languages. The most well known of these lexical databases is perhaps CELEX (Baayen et al., 1995), which contains a large amount of lexical information for English, German, and Dutch. Language-specific lexical databases have also been developed. The MRC psycholinguistic database (Coltheart, 1981), Lexique (New et al., 2001, 2004, 2007), and dlexDB (Heister et al., 2011), for instance, are lexical databases for English, French, and German, respectively.

In addition to these lexical databases, so called "lexicon projects" have recently made their way into the literature. These lexicon projects primarily aim at providing lexical decision reaction times or naming latencies, but typically also supply a substantial number of lexical variables. The English Lexicon Project (ELP), for instance, contains lexical decision and word naming for 40,481 words, but also provides 30 numerical variables and 4 categorical variables for each word (Balota et al., 2007). The British Lexicon Project (BLP) contains 14,365 English words and non-words and their lexical decision times (Keuleers et al., 2012). Lexicon projects have been developed not only for English, but also for other languages. The French Lexicon Project (FLP) provides lexical decision data for 38,940 French words and non-words (Ferrand et al., 2010). The Dutch Lexicon Project (DLP) is a database of lexical decision times for more than 14,000 Dutch words (Keuleers, Diependaele, & Brysbaert, 2010) and was recently extended to 30,000 words (Brysbaert et al.,

2016). Naming latencies for Italian are available in Barca et al. (2002). Finally, Yap et al. (2010) constructed a lexicon project for a less-studied language: the Malay Lexicon Project (MLP). The MLP is a lexical database that contains 11 numerical and 2 categorical variables for 9,592 Malay words, as well as lexical decision and naming latencies for a subset of 1,510 of these words.

Compared to English, far fewer lexical resources exist for Mandarin Chinese. Nonetheless, databases with lexical information exist for Chinese as well. For traditional Chinese, Taiwan Sinica has recently compiled a large-scale lexical database (Y. N. Chang et al., 2016), with naming latencies and 12 numerical variables for 3,314 characters. The naming latencies in this database are based on 20 naming latencies per item. For simplified Chinese, three lexical resources have recently been developed. The first lexical database made available for Chinese by Y. Liu et al. (2007) contains word naming latencies and 15 lexical predictors for 2,423 one-character words in simplified Chinese. The Chinese Lexicon Project (CLP) provides lexical decision latencies for 2,500 characters in simplified Chinese and is released without lexical variables (Sze et al., 2014). Finally, SUBTLEX-CH (Cai & Brysbaert, 2010) is a collection of character and word frequency counts based on movie subtitles in simplified Chinese.

Here, we present a new large-scale lexical database for simplified Chinese: the Chinese Lexical Database (CLD). The CLD contains lexical information for 5,242 unique characters and 30,645 one-character words and two-character words. The database contains 141 numerical variables and 23 categorical variables. The CLD database is available for download and an online interface with basic search functionality is provided at `http://www.chineselexicaldatabase.com`.[1] First, we describe how words were selected for inclusion in the CLD. Next, we describe the categorical variables and the numerical variables in the CLD. Finally, we briefly introduce the online interface to the database.

[1] Access to `http://www.chineselexicaldatabase.com` is password-protected until this dissertation is published. The password is `75090246`.

2.2 Word selection

We created a list of words that were present in the SUBTLEX-CH word frequency list (Cai & Brysbaert, 2010), as well as in the Contemporary Chinese Dictionary (Xiandai Hanyu Cidian, Chinese Academy of Social Sciences, 2012) and for which both characters appeared in the Chinese Character Dictionary that is available online at `http://www.mandarintools.com/chardict.html`. Two-character words that were a repetition of the same character were not included.

The original word list consisted of 30,801 one-character words and two-character words. We removed 131 words with traditional Chinese characters that have a simplified version, 14 proper nouns, and 10 Japanese Hanzi (Kanji). The final list therefore contains 30,645 one-character words and two-character words. The number of unique characters in this list is 5,242.

Not all characters can be used as independent words: 532 of the 5,242 unique characters do not appear as an independent word in the CLD. As a result, the CLD consists of 4,710 one-character words and 25,935 two-character words. This makes the CLD the largest lexical database that is available for simplified Chinese.

2.3 Categorical variables

The CLD consists of 23 categorical variables that can be grouped into 10 classes. Class 1 contains the orthographic forms of the word and its characters (`Word`, `Character 1`, `Character 2`). As mentioned above, the CLD consists of 4,710 one-character words and 25,935 two-character words. For the 4,710 single character words, `Character 2` is set to "NA".

The lexical variables in Class 2 contain the Pinyin for the word and its characters (`Pinyin`, `Character 1 Pinyin`, `Character 2 Pinyin`). Pinyin literally means "spell the sound" and translates Chinese characters into a romanized form based on their pronunciations. The Pinyin annotation in the CLD is based on a publicly available Pinyin annotator developed by Xiao (2010-2015). Although the orthographic form of Chinese characters does not provide tonal information (see below), numbers or diacritics can be used in Pinyin to indicate the tone associated with a word or character (e.g., "ma2" or "má"). In the CLD, we represent tonal information in the numeric form (e.g., "ma2").

Class 3 provides phonetic information about the characters and the words in IPA format. IPA transcriptions are provided both for the word and for its characters (`IPA`, `Character 1 IPA`, `Character 2 IPA`). IPA transcriptions were obtained through the application of a set of 37 Pinyin to IPA conversion rules to the Pinyin variables described above. The set of conversion rules is based on a subset of the Pinyin to IPA conversion rules on Wikipedia (Wikipedia, 2016).

The variables in Class 4 encode tonal information for both characters (`Character 1 Tone`, `Character 2 Tone`). In Mandarin Chinese, the number of unique syllables is limited. This would lead to overwhelming amounts of homophony if pronunciations were purely based on syllables. The use of tones helps overcome this problem and allows for the phonological differentiation of characters with the same syllabic structure. Mandarin Chinese has 5 tones: 4 main tones and a neutral tone. The five tones have distinctive pitch contours, which are high-level (tone 1), high-rising (tone 2), low/dipping (tone 3), high-falling (tone 4) and neutral (tone 5).

The number of unique character-tone combinations is greater than the number of unique characters in the CLD. For instance, the character 兴 is pronounced as "[ɕiŋ4]" in the word 高兴 ("happy") and as "[ɕiŋ1]" in the word 兴奋 ("excited"). In total, there are 5,685 unique character-tone combinations for the 5,242 unique characters in the CLD. The distribution of tones across the 5,685 unique character-tone combinations in the CLD is presented in Table 2.1. The four main tones have relatively similar frequencies, with tone 4 being somewhat more frequent (31.89%) and tone 3 being somewhat less frequent (15.97%) than tones 1 (24.61%) and 2 (24.24%). Tone 5, also known as the "neutral tone" is the least frequent tone (3.29%).

Table 2.1: Distribution of tone across unique character-tone combinations

	count	percentage
Tone 1	1399	24.61%
Tone 2	1378	24.24%
Tone 3	908	15.97%
Tone 4	1813	31.89%
Tone 5	187	3.29%

Information about the structure of characters is encoded in the categorical variables in Class 5 (`Character 1 Structure`, `Character 2 Structure`). We divided the characters into six different structures: Left-Right (e.g., 明 "brightness" or "tomorrow"), Left-Right-Bottom (e.g., 边 "side"), Up-Down (e.g., 草 "grass"), Circle (e.g., 回 "return"), Half Circle (e.g., 区 "area"), and Single (e.g., 开 "open"). The distribution of structures across the 5,242 unique characters is illustrated in Table 2.2. The most common structures are Left-Right (62.42%) and Up-Down (23.62%), with 86.04% of all characters having one of these two structures.

Table 2.2: Distribution of character structure across unique characters

	count	percentage
Left-Right	3272	62.42%
Left-Right-Bottom	121	2.31%
Up-Down	1238	23.62%
Circle	23	0.44%
Half-Circle	277	5.28%
Single	311	5.93%

The sixth class of categorical variables in the CLD describes the type of a character (`Character 1 Type`, `Character 2 Type`). As noted by Hsieh (2006), Chinese characters can be divided into six different types, which are called the "Six Writings" in the Chinese linguistic literature (c.f., Yip, 2000). According to Hsieh (2006), there are 4 basic types of character construction among these 6 types: pictographic, pictologic, pictosynthetic, and pictophonetic. The other 2 types are "phonetic loan character" and "cognate", which according to Hsieh (2006) are extensions of the 4 basic types that describe ways of using characters.

We restricted character types in the CLD to the 4 basic types of character construction. Pictographic characters were the earliest type of character in Chinese and originate from non-linguistic symbolic systems. These types of characters resemble the shapes of objects. The physical forms of the characters 川 ("river") and 山 ("mountain"), for instance, resemble the objects denoted by them.

The second type of characters are pictologic characters. Pictologic characters refer to objects that do not have a concrete, easy to depict shape. Typically, although not always, a stroke is added to a pictographic character in order to refer to more specific or more abstract concepts. Consider for instance the character 刃 ("blade"),

which was derived from the pictographic character 刀 ("knife"). The extra stroke on the left side of the character 刃 ("blade") represents the blade of a knife.

As the writing system evolved, pictosynthetic characters came into existence. Pictosynthetic characters consist of multiple pictographic characters that are combined to form a new character. In the most basic form of a pictosynthetic character, the repetition of a pictographic character forms a new character. This is, for instance, the case for 木 ("wood"), which both through duplication 林 ("trees") and through triplication 森 ("forest") forms new characters. The combination of two different pictographs can also create a new character: combining 日 ("sun") and 月 ("moon") results in the new character 明 ("brightness"), which describes a shared semantic property of both characters.

The fourth character type is pictophonetic. Pictophonetic characters are based on a combination of a semantic radical and a phonetic radical (see the description of Class 7 and Class 8 below for more information about semantic and phonetic radicals). The character 清 ("to clean") is an example. It consists of the semantic radical 氵 ("liquid", "[ʂui3]") and the phonetic radical 青 ("[t^hɕiŋ1]"). The phonetic radical determines the pronunciation of the character 清, which is "[t^hɕiŋ1]".

Finally, there are characters that cannot be categorized as one of the four main character types. Character type is encoded as "Other" for these characters. The most common type of characters encoded as "other" are characters that were simplified to the extent that they no longer fall into one of the 4 main types. For instance, the traditional Chinese character 廣 ("broad"; "[kuɑŋ3]"), which is a pictophonetic character, was simplified into 广 ("broad", "[kuɑŋ3]"). The phonetic radical 黃 ("[xuɑŋ2]") of the original character 廣 was removed entirely in the simplified character 广. The semantic radical remained and is used as an independent character.

The distribution of character type across unique characters is shown in Table 2.3. Pictophonetic characters (72.55%) are the most frequent character type by a large margin. The next most frequent character types are pictosynthetic (13.75%) and pictographic (4.85%). Only 0.93% of all characters was pictologic. Less than a tenth of all characters (7.92%) did not fall into one of the 4 basic character construction types and was categorized as "Other".

The variables in Class 7 provide information about the semantic radicals of the characters (`Character 1 SR`, `Character 2 SR`). When looking up words in a Chinese dictionary, radicals (known as 部首, or "section-head", in Chinese) serve as a search

Table 2.3: Distribution of character type across unique characters

	count	percentage
Pictograph	254	4.85%
Pictologic	49	0.93%
Pictosynthetic	721	13.75%
Pictophonetic	3803	72.55%
Other	415	7.92%

index. The radicals used to look up words in a dictionary are sometimes called 义符 ("meaning-symbol"), or semantic radicals. Semantic radicals are typically associated with a semantic concept. The semantic radical 氵, for instance, appears in words like 河 ("river"), 海 ("ocean"), or 泪 ("tears"), all 3 of which have meanings related to the semantic concept "liquid".

Although most radicals used for looking up words in a dictionary carry semantic information related to the meaning of the character they appear in, this is not always the case. Radicals that are used for dictionary lookup, but that do not have a systematic semantic relationship with the words they appear in are called 形体-部首 ("shape-radical"). The word 东 ("east") is an example of this. The radical that is used to look up this word in a dictionary is 一 ("one"), which is semantically independent from the meaning of the word 东 ("east"). Although there is no direct semantic relationship between these "shape radicals" and the words they appear in, we qualify these radicals as semantic radicals in the CLD.

The semantic radicals in the CLD were obtained through the Chinese Character Dictionary (`http://www.mandarintools.com/chardict.html`). In total, the 5,242 unique characters in the CLD contain 250 unique semantic radicals. Consistent with the observations of Feldman and Siok (1999a), there is considerable variation with respect to the number of characters in which a semantic radical is used. The most frequent semantic radicals in the CLD are 口 ("mouth"; appears in 286 unique characters), 扌 ("hand"; 255 characters), 艹 ("grass"; 244 characters), 木 ("wood"; 225 characters), and 亻 ("person"; 222 characters).

Whereas the lexical variables in Class 7 encode information about the *semantic* radicals in both characters of a word, the variables in Class 8 provide the phonetic radicals for both characters (`Character 1 PR`, `Character 2 PR`). Generally speaking, phonetic radicals carry information about the pronunciation of a character. This information, however, is not always reliable. While sometimes the pronunciation

of the phonetic radical is identical to the pronunciation of the character (e.g., 榆 ("elm"), as well as its phonetic radical 俞 are pronounced as "[iu2]"), this is not necessarily the case. For the character 翅 ("wing"), for instance, the character pronunciation is "[t^hʂi4]", whereas the pronunciation of its phonetic radical 支 is "[tʂi1]". The character pronunciation therefore is similar, but not identical to the pronunciation of the phonetic radical. For the character 拓, the pronunciation of the character ("[t^huɔ4]") and its phonetic radical 石 ("[ʂi2]") are entirely independent.

The pronunciation of a phonetic radical in isolation is represented by the categorical variables in Class 9 (`Character 1 PR Pinyin`, `Character 2 PR Pinyin`). Most phonetic radicals have a unique pronunciation when presented in isolation. Some phonetic radicals, however, have multiple possible pronunciations. For these phonetic radicals, we based our choice of the pronunciation on the character and word context. For example, the character 且, is associated with 2 Pinyin forms: "ju1" and "qie3" and is used as a phonetic radical in a number of pictophonetic characters. The pronunciation of the characters 沮 ("ju3") and 阻 ("zu3") are based on the form "ju1", whereas the pronunciation of the character 姐 ("jie3") is based on the form "qie3". For the characters 阻 and 沮 we therefore encoded the phonetic radical Pinyin as "ju1", whereas for the character 姐, we set phonetic radical Pinyin to "qie3".

Whereas all characters contain a semantic radical, not all characters contain a phonetic radical. Of the 5,242 unique characters in the CLD, 3,796 have a phonetic radical (72.42%). The most frequent phonetic radicals in the CLD are 非 ("[feɪ1]"; appears in 19 unique characters), 各 ("[kɤ4]"; 19 characters), 隹 ("[tʂui1]"; 18 characters), 包 ("[pɑʊ1]"; 18 characters), and 且 ("[tɕu1]" or "[t^hɕiɛ3]"; 18 characters).

The variables in the final class of categorical variables, Class 10, indicate whether the pronunciation of a phonetic radical is regular (`Character 1 PR Regularity`, `Character 2 PR Regularity`). As mentioned above, the pronunciations of a phonetic radical and the character it occurs in may or may not be the same. Whenever the pronunciation of a character is identical to the pronunciation of its phonetic radical, phonetic radical regularity is set to 1 (e.g, 榆 ("elm"), as well as its phonetic radical 俞 are pronounced as "[iu2]"; `Character 1 PR Regularity` for the single character word 俞 is thus set to 1).

When the pronunciation of a character is not identical to the pronunciation of the phonetic radical, phonetic radical regularity is set to 0. For the character 拓, for instance, the pronunciation of the character ("[t^huɔ4]") and its phonetic radical 石

("[ṣi2]") are different. `Character 1 PR Regularity` for the single character word 拓 is therefore set to 0. We used a strict definition of regularity, which considers not only differences in phonemes, but also differences in tones. The pronunciation of the single-character word 案 is "[än4]", whereas the pronunciation of its phonetic radical 安 is "[än1]". Given the difference in tone, `Character 1 PR Regularity` is set to 0 for the word 案.

In total, there are 4,109 unique combinations of a character, a phonetic radical, a character pronunciation and a phonetic radical pronunciation. For 1,404 (34.17%) of these combinations, the pronunciation of the character is identical to the pronunciation of the phonetic radical, whereas for 2,705 (65.83%) of these combinations it is different. By comparison, Fan et al. (1984) found that the pronunciation of the phonetic radical was identical to the pronunciation of the character for 26.3% of the 5,990 characters under their investigation, whereas Zhou and Marslen-Wilson (1999) reported that "less than 30% of the complex characters composed of semantic and phonetic radicals have exactly the same pronunciations as their phonetics radicals".

The 23 categorical variables in the CLD provide categorical information about the orthographic and phonological properties of a word and its characters, as well as about the type, the structure and the phonetic and semantic radicals of a character. In the next section, we discuss the numerical variables in the CLD.

2.4 Numerical variables

2.4.1 Clustering

The CLD consists of 141 numerical variables. These numerical variables are not mutually independent. There are strong correlations between subsets of the full set of numerical variables. Rather than describing the numerical variables independently, we therefore describe the set of numerical variables on the basis of the results of a clustering analysis on the full set of numerical variables.

The clustering analysis is based on a self-organizing map (SOM; Kohonen, 1982). SOMs are artificial neural networks that are trained in an unsupervised manner on the basis of a competitive learning algorithm (c.f., Hebbian learning; Hebb, 1949) with a forgetting term (see Haykin, 1998). Unlike the output neurons in traditional neural networks, the output neurons in a SOM are arranged in a low-

dimensional (most commonly 1D or 2D) space. Neurons that are closer together in this topological space encode more similar information than neurons that are farther apart. SOMs thus help organize complex data sets with a high dimensionality into an easily interpretable low-dimensional map.

For each output neuron the model constructs a vector of weights between all input neurons and that output neuron. The initial weight vectors between input units and outcome units are set randomly. When the first input pattern is presented to the model, the best matching unit (BMU) is identified. The BMU is the output neuron for which the summed absolute distance between the input pattern and the weight vector is minimal. The weight vector for the BMU is updated based on the difference between the input pattern and the weight vector. Furthermore, the weight vectors for other output neurons are updated based on the distance between that output neuron and the BMU in the SOM. The weight vectors for output neurons that are closer to the BMU are updated more than the weight vectors for output units that are farther from the BMU. The extent to which weight vectors are updated (i.e., the learning rate) depends not only on the proximity to the BMU, but also decreases as a function of time. The process of updating weights is repeated iteratively until the model stabilizes.

Typically, SOM maps are fitted to data sets in which the rows are observations and the columns are variables. The CLD lexical database, however, contains a great amount of missing data. Most strikingly, all predictor values related to character 2 are missing for 1 character words. Another source of missing data are phonetic radical measures: 1,446 of the $5,242$ unique characters do not have a phonetic radical (27.58%). In their basic form, self-organizing maps do not allow for missing data.

Implementations of SOMs in which missing data are allowed are available. In these implementations, missing values are most commonly handled by computing the BMU on the basis of the output neuron weight vectors for which the corresponding value(s) is/are not missing in the input vector. Given the number of missing values in the CLD, however, this would result in considerably different evaluation criteria for different words. We therefore decided to use the squared 140 by 140 correlation matrix for the numerical variables in the CLD, rather than the raw data set, as the input data for SOM used here. To prevent the correlation matrix from being overly influenced by predictor outlier values or non-symmetrical distributions (see Baayen, 2008), we used the Spearman correlation matrix, rather than the Pearson correlation matrix. Correlations were based on pairwise complete observations. The numerical

predictor `Length` (the length of the word in characters; values: 1 (4,710 words) , 2 (25,935 words)) was not included in the correlation matrix, as correlations between `Length` and numerical predictors describing properties of the second character are, by definition, not available.

We fitted a SOM to the correlation matrix using the `kohonen` package for the statistical software R (version 2.0.19; Wehrens & Buydens, 2007). Successful clustering on a SOM requires a sufficient number of output nodes. To allow for optimal clustering performance, we therefore used a 10 by 10 hexagonal grid of output neurons despite the low counts (i.e., the number of "observations" for which the input vector is most similar to the weight vector of a given output neuron) that resulted from this approach (mean count: $\frac{140}{10*10} = 1.40$). The learning rate parameter linearly decreased from 0.05 to 0.01 as a function of time. The correlation matrix was presented to the model 200 times. As can be seen in Figure 2.1, 200 iterations were sufficient, with the mean distance of an input vector to the weight vector of the BMU starting to stabilize after about 100 presentations of the correlation matrix.

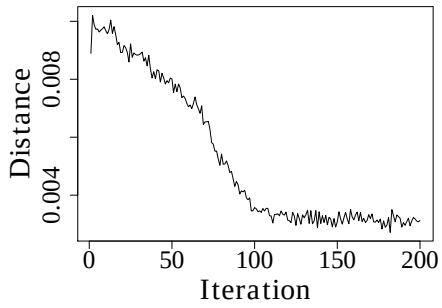

Figure 2.1: Kohonen SOM training. Change in mean distance to closest unit as a function of training iteration.

The `kohonen` package provides a measure of topographical error (i.e., map smoothness) that is based on the average distance on the SOM between pairs of output neurons with the most similar weight vectors. The SOM reported here had a topographical error of 1.059. The mean topographical error for $1,000$ similar SOMs with a different random initialization of weights was 1.132, with a standard deviation of 0.040. Out of the 1000 similar SOMs, only 11 SOMs (1.10%) had a topographical error smaller than or equal to the SOM presented here. The SOM reported here is therefore characterized by limited topographical error, which indicates that the map quality is satisfactory.

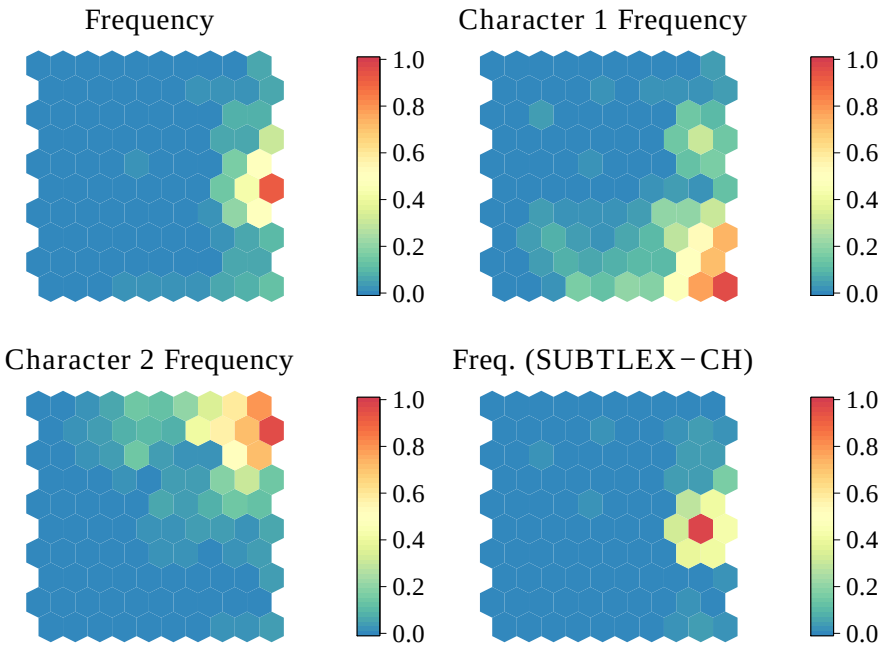

Figure 2.2: Predictor heatmaps for Kohonen SOM. Top left panel: Frequency. Top right panel: Character 1 Frequency. Bottom left panel: Character 2 Frequency. Bottom right panel: Frequency (SUBTLEX-CH).

The output of a SOM is a matrix of weights, with the output neurons as rows and the columns of the input data (i.e., the numerical predictors) as columns. The distribution of a given predictor across the SOM can be visualized through a heatmap of the distribution of weights for that predictor across the map. Figure 2.2 shows the topographical distribution of 4 numerical variables in the CLD. The first three predictors are `Word Frequency`, `Character 1 Frequency`, and `Character 2 Frequency`. These are the word and character frequencies derived from a corpus of web pages described in the discussion of frequency measures below. The fourth predictor is `Word Frequency SUBTLEX-CH`, which is the frequency of the word in the SUBTLEX-CH corpus (Cai & Brysbaert, 2010).

Figure 2.2 indicates that all 4 frequency measures are located at the right side of the map. The topographical distribution of word frequency in the corpus of web pages and the SUBTLEX-CH is highly similar, with high weights near the middle-right of the map. The output neurons with the highest weights for the character frequencies are topologically close to the "word frequency neurons", with high weights for `Character 1 Frequency` in the bottom right of the SOM and high weights for `Character 2 Frequency` in the top right of the SOM.

Heatmaps are highly informative about the topological distribution of a given predictor across the SOM. For the current purpose of grouping similar numerical variables together, however, it is important to know which predictors have similar topological distributions. One way to look at this is to use a Euclidean distance matrix of the weight matrix of the SOM as input to a clustering algorithm. Here, we applied the agglomerative clustering algorithm in the `hclust` function in R (with complete linkage) to this distance matrix. We limited the number of clusters to 21 based on a subjective assessment of the resulting clusters in the context of describing different groups of numerical predictors.

Figure 2.3 shows the results of the agglomerative clustering technique on (a Euclidean distance matrix of) the weight matrix of the SOM. We assigned the 140 numerical variables that served as input for the SOM to clusters based on the cluster membership of the output neuron with the maximum weight for a given lexical variable. Each cluster represents a group of lexical variables with a similar topology on the SOM. We added cluster numbering and colour coding of clusters to Figure 2.3 for ease of interpretation.

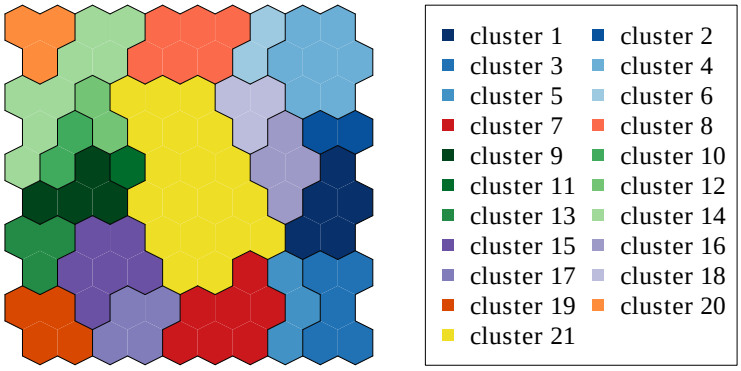

Figure 2.3: Clusters in Kohonen SOM. Cluster numbering and colour coding were manually added for ease of interpretation.

The colour coding in Figure 2.3 distinguishes 6 groups of clusters. Blue clusters (Group 1, Cluster 1 to Cluster 6) contain frequency measures. Red clusters (Group 2, Cluster 7 and Cluster 8) consist of lexical variables that are related to the visual complexity of a variable. Lexical predictors that describe phonological properties of a character or word are in green clusters (Group 3, Cluster 9 to Cluster 14). Purple clusters (Group 4, Cluster 15 to Cluster 18) contain information about homographs, while the numerical variables in orange clusters (Group 5, Cluster 19 and Cluster 20)

encode information about homophones. Finally, the yellow cluster (Group 6, Cluster 21) contains a variety of variables that did not fall into any of the other clusters. As can be seen in Figure 2.3, the topological organization of the SOM is sensible, with (groups of) clusters that contain conceptually similar numerical variables occupying neighbouring areas of the map. Consistent with Figure 2.2, frequency measures are grouped together at the right of the map. The red, purple, orange and green clusters for lexical variables related to the first character of a word are in the lower part of the map (darker shades), whereas the corresponding clusters for the second character of a word are in the upper part of the map (lighter shades).

For clarity, we should note that the SOM presented here was semi-automatically selected from a series of SOMs, based on three criteria. First, we considered only those SOMs with a topographical error below 1.06 (approximately 0.12% of all SOMs). Second, we inspected candidate SOMs based on the interpretability and neatness of the clusters in the agglomerative clustering technique described above. Third, we rejected SOMs with clusters that are spatially non-continuous. As demonstrated above, this procedure resulted in a quantitatively and qualitatively satisfactory SOM. We use this SOM to organize the discussion of the numerical variables in the CLD.

2.4.2 Group 1: frequency measures

As shown in Figure 2.3 there are 6 clusters in different shades of blue. Each of these 6 clusters contains a different type of frequency-related numerical variables. Table 2.4 provides an overview of the 6 clusters in Group 1. The types of numerical variables in these clusters are related to word frequency (Cluster 1), between-character associations (Cluster 2), character 1 frequency (Cluster 3), character 2 frequency (Cluster 4), character 1 entropy (Cluster 5) and character 2 entropy (Cluster 6). Below, we describe each of these clusters in more detail.

2.4.2.1 Cluster 1: word frequency

As can be seen in Table 2.4, Cluster 1 consists of 6 numerical variables: 3 word frequency (per million) and 3 contextual diversity measures. These measures are from 3 different resources: the Simplified Chinese Corpus of Webpages (henceforth SCCoW; Shaoul et al., 2016), the Gigaword corpus (Graff & Chen, 2003), and SUBTLEX-CH (Cai & Brysbaert, 2010). In this section, we first describe each of these three resources and then compare the frequency and contextual diversity measures from the SCCoW, the Gigaword corpus and SUBTLEX-CH. We demonstrate

that the measures derived from the SCCoW provide most explanatory power for a set of naming latencies for the 30,645 words in the CLD.

The SCCoW (Shaoul et al., 2016) is a new corpus of simplified Chinese. Using a web crawling robot, we collected text from over 40 million web pages that were located in the ".cn" top level domain or that appeared within a ".com" domain and contained only simplified Chinese text. The methodology for cleaning the corpus

Table 2.4: Overview of numerical predictors: frequency measures (Group 1)

	mean	median	sd	min	max	NA
Cluster 1: word frequency						
Frequency (SCCoW)	22.68	1.38	305.49	0.00	45362.39	434
Frequency (Gigaword)	22.20	1.29	292.64	0.00	43115.16	312
Frequency (SUBTL)	30.28	0.80	587.42	0.03	50155.13	0
CD (SCCoW)	0.77	0.08	3.31	0.00	99.59	434
CD (Gigaword)	0.67	0.06	3.14	0.00	99.12	312
CD (SUBTL)	3.25	0.34	10.67	0.02	100.00	0
Cluster 2: association measures						
PMI	3.84	3.62	4.54	-10.47	25.88	4748
Position-specific PMI	5.84	5.79	4.49	-9.02	25.88	4748
T-Score	27.91	4.12	73.46	-8.30	630.97	4748
Cluster 3: character 1 frequency						
C1 Freq. (SCCoW)	688.16	206.00	1193.56	0.00	28187.84	36
C1 Freq. (Gigaword)	694.38	208.77	1243.93	0.00	26364.54	122
C1 Freq. (SUBTL)	676.59	153.24	1878.62	0.02	43956.70	0
C1 CD (SCCoW)	22.20	11.78	24.34	0.00	99.60	36
C1 CD (Gigaword)	20.51	10.21	23.86	0.00	99.12	122
C1 CD (SUBTL)	48.73	46.13	35.83	0.02	100.00	0
C1 Family Size	37.61	25	39.59	0	436	0
C1 Family Freq.	662.04	182.05	1039.03	0.00	6221.84	0
C1 Friends	35.81	24	37.68	0	365	0
C1 Friends Freq.	799.46	192.99	1452.13	0.00	45580.04	0
C1 PR Friends	29.63	22	27.39	0	143	14478
C1 PR Friends Freq.	472.25	123.64	947.04	0.00	45580.04	14478

Table 2.4 (continued)

	mean	median	sd	min	max	NA
Cluster 4: character 2 frequency						
C2 Freq. (SCCoW)	783.67	360.85	1179.17	0.01	28187.84	4710
C2 Freq. (Gigaword)	760.61	359.15	1183.14	0.00	26364.54	4710
C2 Freq. (SUBTL)	762.00	240.13	1818.75	0.02	43956.70	4710
C2 CD (SCCoW)	25.65	18.36	23.43	0.00	99.60	4710
C2 CD (Gigaword)	23.18	15.48	22.57	0.00	99.12	4710
C2 CD (SUBTL)	56.60	59.59	33.86	0.02	100.00	4710
C2 Family Size	51.86	33	64.82	1	436	4710
C2 Family Freq.	836.48	352.89	1136.42	0.00	6221.84	4710
C2 Friends	47.65	31	55.35	0	365	4710
C2 Friends Freq.	927.32	367.96	1470.30	0.00	45712.63	4710
C2 PR Friends	36.99	29	30.27	0	143	17944
C2 PR Friends Freq.	646.46	248.09	1166.32	0.00	45712.63	17944
Cluster 5: character 1 entropy						
C1 Entropy	2.39	2.46	1.16	0.00	5.03	4713
C1 Trigram Entropy	9.88	10.17	2.04	0.00	17.54	36
Cluster 6: character 2 entropy						
C2 Entropy	2.54	2.56	1.27	0.00	5.74	4711
C2 Trigram Entropy	10.17	10.34	1.68	0.29	17.54	4710
C2 PR Frequency	1182.04	654.57	2064.32	0.03	28622.72	17944

was adapted from Baroni et al. (2009). The corpus was cleaned in the following manner: only files of mime type text/html that were between 5 and 1000 KB in size were retained. Custom code was used to remove identical documents and all non-text regions (HTML, CSS and javascript) as well as 'boilerplate' text (navigation buttons, disclaimers, copyright statements and the like).

One of the goals in creating the SCCoW was the detection and removal of documents written using traditional Chinese characters or in a non-Chinese language, such as English. Custom code was used to calculate the proportion of characters in a document that are never used in simplified Chinese (certain characters are used in both the traditional and simplified Chinese writing systems) or were non-Chinese. Following a visual inspection of the output of this simplified Chinese document detector, we set the rejection criterion at 10% (i.e., any document with 10% or more

characters in traditional Chinese or non-Chinese was rejected from the corpus). The cleaned corpus was segmentized using the conditional random field Chinese Word Segmenter developed by P. C. Chang et al. (2008).

After cleaning and segmentation, the SCCoW is 2.7 GB in size and consists of 607,290 documents. A reading of a random sampling of texts from the corpus revealed that it contained a mix of encyclopedic, narrative, journalistic, bureaucratic and conversational registers. According to a Unix word count excluding whitespace, but including punctuation marks, traditional Chinese characters and words, and non-Chinese characters and words, the corpus contains 466,551,657 words and 773,697,216 characters. The inclusion of all Chinese and non-Chinese characters and punctuation marks in the frequency counts comes with the advantage of having frequency counts available for all linguistic elements that appear in web pages of simplified Chinese.

We would like to point out that the inclusion of all Chinese and non-Chinese characters and punctuation marks in the frequency counts leads to somewhat deflated frequency counts per million characters or per million words as compared to corpora that provide frequency counts per million on versions of corpora that exclude these elements. To allow for a recalculation of frequency per million excluding certain elements, we provide raw frequency counts in the downloadable version of the CLD, as well as through the online search interface.

The second corpus from which we obtained frequency measures, the Gigaword corpus (Graff & Chen, 2003), is a corpus of newswire text data from the Central News Agency of Taiwan (CNA) and the Xinhua News Agency of Beijing (XIN). In contrast to China, Taiwan uses the traditional Chinese writing system. A certain proportion of the texts in the Gigaword corpus, hence, is written in traditional Chinese. Nonetheless, we decided to supply frequency measures for the Gigaword corpus, because it is much larger than other existing corpora of Chinese. Again, we segmentized the Gigaword corpus using the Chinese Word Segmenter (P. C. Chang et al., 2008). After segmentation the corpus comprises 4.06 GB and contains 1,228,381 documents. These documents contain a total of 718,545,994 words and 1,200,562,229 characters. As before, these counts exclude whitespace, but include punctuation marks, traditional Chinese characters and words, and non-Chinese characters and words.

The third corpus, SUBTLEX-CH (Cai & Brysbaert, 2010), is a corpus in a series of film and television subtitle corpora for different languages, including British (Van Heuven et al., 2014) and American English (Brysbaert & New, 2009), German (Brysbaert et al., 2011), Dutch (Keuleers, Brysbaert, & New, 2010), Spanish (Cuetos et al., 2011), Greek (Dimitropoulou et al., 2010) and Polish (Mandera et al., 2015). The SUBTLEX-CH corpus consists of 46.8 million characters and 33 million words. The character and word frequency lists are publicly available.

Cai and Brysbaert (2010) compared the SUBTLEX-CH frequency and contextual diversity measures to similar measures computed for other existing corpora, such as the Language Corpus System of Modern Chinese Study (LCSMCS, H. L. Sun et al., n.d.), the Center of Chinese Linguistics corpus (CCL, Center for Chinese Linguistics, 2006), and the Lancaster Corpus of Mandarin Chinese (LCMC, McEnery & Xiao, 2003). They demonstrated that the SUBTLEX-CH frequency and contextual diversity measures were superior predictors for lexical decision and word naming latencies as compared to corresponding measures from the other corpora.

For all three corpora describe above, we extracted word frequency measures (`Frequency (SCCoW)`, `Frequency (Gigaword)`, `Frequency (SUBTLEX-CH)`), as well as contextual diversity measures (`CD (SCCoW)`, `CD (Gigaword)`, `CD (SUBTLEX-CH)`). Contextual diversity is defined as the percentage of documents in which a character or word occurs. In the SCCoW, for instance, the word 不 ("no" or "not") occurs in $406,786$ of all $607,290$ web pages, for a contextual diversity of 66.98%. Together these 6 numerical variables form Cluster 1, the cluster of word frequency measures.

Table 2.4 shows a number of descriptive statistics for each of the 6 predictors in Cluster 1. For each numerical variable the mean, median, standard deviation (sd), minimum (min), maximum (max) and the number of missing data points (*NA*) are provided. Since the list of words was selected partly based on presence in SUBTLEX-CH, 0 words are missing in the SUBTLEX-CH corpus. However, 434 words are not in the SCCoW and 312 words are not in the Gigaword corpus. We set the SCCoW and Gigaword frequency for these words to *NA* (as opposed to the alternative 0), because in some cases the absence of a word in these corpora is due to suboptimal segmentation rather than to true absence in the corpus. Table 2.4 also shows that the means are much larger than the medians for all word frequency measures. This is the case, because – as is common across languages – frequency measures in Chinese follow a Zipfian distribution (Zipf, 1932).

Table 2.5: Pairwise Spearman correlations for the numerical variables in Cluster 1. Abbreviations: FSC = Frequency (SCCoW), FGI = Frequency (Gigaword), FSU = Frequency (SUBTLEX-CH), CDSC = CD (SCCoW), CDGI = CD (Gigaword), CDSU = CD (SUBTLEX-CH).

predictor	FSC	FGI	FSU	CDSC	CDGI	CDSU
FSC	-					
FGI	0.924	-				
FSU	0.641	0.661	-			
CDSC	0.993	0.922	0.645	-		
CDGI	0.919	0.994	0.667	0.927	-	
CDSU	0.661	0.674	0.988	0.670	0.684	-

Table 2.5 shows the pairwise correlations between the numerical variables in Cluster 1. As can be seen in Table 2.5, all pairwise correlations between the 6 predictors are positive. All correlations in Table 2.5 are significant at an α level of 0.001. This indicates that the word frequency cluster is homogeneous. Furthermore, Table 2.5 shows that the word frequency ($r = 0.924$) and contextual diversity ($r = 0.927$) measures from the SCCoW and Gigaword corpora are highly similar. This suggests that there is little difference between the language used in web pages and the language used in newswire texts.

This concludes our discussion of measures in the word frequency cluster. The clusters that follow, however, contain many numerical variables that are calculated on the basis of frequency counts. To prevent an unwieldy number of numerical variables, we decided to calculate these measures on the basis of frequency measures for one source only. To determine which source to use, we compared the performance of the frequency and contextual diversity measures from each of the 3 corpora described above in accounting for the naming latencies from a native reader of simplified Chinese for each of the $30,645$ words in the CLD (see Chapter 3).

The correlations of the frequency and contextual diversity measures extracted from the 3 corpora with the observed naming latencies are presented in Table 2.6. Table 2.6 includes not only word-level predictors, but also character-level frequency and contextual diversity measures. These character-level predictors will be introduced in more detail in our discussion of Cluster 3 and Cluster 4.

Table 2.6: Correlations of (logged) frequency and contextual diversity measures from the SCCoW, the Gigaword corpus and SUBTLEX-CH with observed naming latencies for a native reader of simplified Chinese. Maximum correlations for each predictor are printed in bold font.

predictor	SCCoW	Gigaword	SUBTLEX-CH
Word Frequency	**-0.232**	-0.231	-0.219
Character 1 Frequency	**-0.388**	-0.385	-0.387
Character 2 Frequency	**-0.227**	**-0.227**	-0.217
Word CD	**-0.235**	-0.233	-0.224
Character 1 CD	**-0.389**	-0.386	-0.364
Character 2 CD	-0.223	**-0.224**	-0.211

Table 2.6 shows that correlations with the observed naming latencies are similar for frequency and contextual diversity measures from the 3 corpora. The average correlation of the 6 frequency and contextual diversity measures with observed naming latencies in the SCCoW is -0.282, as compared to -0.281 in the Gigaword corpus and -0.270 in SUBTLEX-CH.

To determine which corpus provided frequency and contextual diversity measures that best accounted for the observed naming latencies, we carried out a series of paired t-tests on the correlations listed in Table 2.6. These t-tests indicated that correlations with observed naming latencies are weaker for the SUBTLEX-CH measures than for the SCCoW ($t(5) = -3.842$, $p = 0.012$) and Gigaword measures ($t(5) = -3.517$, $p = 0.017$), although the latter t-test was only marginally significant. The SCCoW and Gigaword measures show very similar correlations with observed naming latencies, which is unsurprising given the average correlation between the SCCoW and Gigaword measures themselves ($r = 0.971$), which is much higher than the average correlation between the SCCoW and SUBTLEX-CH measures ($r = 0.799$) and the average correlation between the Gigaword and SUBTLEX-CH measures ($r = 0.792$). The difference between the correlations with observed naming latencies for the SCCoW measures and the correlations with observed naming latencies for the Gigaword measures is not significant ($t(5) = -1.702$, $p = 0.149$).

Table 2.6 shows that correlations with observed naming latencies for measures for the word and the first character are slightly higher for the SCCoW than for the Gigaword corpus. For measures of the second character the SCCoW performs slightly worse than the Gigaword corpus. Given the slight overall edge of the SCCoW

frequency measures over the Gigaword frequency measures, we decided to calculate all frequency-based measures in the CLD on the basis of the SCCoW character and word frequencies. Using the Gigaword frequency measures, however, would have yielded highly similar results.

As noted above, Cai and Brysbaert (2010) demonstrated that frequency and contextual diversity measures from SUBTLEX-CH outperformed similar measures from other existing corpora. Here, we found that frequency and contextual diversity measures obtained from the SCCoW and – to a lesser degree – the Gigaword corpus perform somewhat better than the corresponding SUBTLEX-CH measures. There are at least two potential explanations for the better performance of the SCCoW and the Gigaword measures as compared to the SUBTLEX-CH measures. First, at 33 million words SUBTLEX-CH comprises less than a tenth of the number of words tokens that the SCCoW (466 million words) and the Gigaword corpus (718 million words) contain. Frequencies for low frequency words may therefore be less accurate for SUBTLEX-CH. Second, the subtitles in SUBTLEX-CH are translated from English. Therefore, frequency counts in SUBTLEX-CH may be influenced by cultural and linguistic differences between anglo-saxon countries and China. By contrast, the web pages and newswire texts in the SCCoW and the Gigaword corpus are authentic Chinese texts that are not sensitive to these differences.

2.4.2.2 Cluster 2: association measures

Cluster 2 contains measures regarding the strength of the association between character 1 and character 2 in two-character words. The three measures in this cluster (`PMI`, `Position-specific PMI`, `t-Score`) are based on the observed frequency (in the SCCoW) of a two-character word and the expected frequency for that word (c.f., Gries, 2010). The expected frequency is defined as:

$$\frac{\text{Character 1 Frequency} * \text{Character 2 Frequency}}{\text{Total Frequency}} \tag{2.1}$$

where Total Frequency is the summed frequency of all 2-character words in the CLD.

For example, the observed frequency per million of the word 苹果 ("apple") is 50.49. To calculate the expected frequency, the frequency of two-character words that contain character 1 (苹, frequency: 50.49)[2] and the frequency of two-character words that contain character 2 (果, 1,587.82), as well as the total frequency of

[2] 苹果 ("apple") is the only two-character word in which 苹 is the first character.

two-character words in the CLD (398,118.96) are required. These numbers give an expected frequency of $\frac{50.49*1{,}587.82}{398{,}118.96} = 0.20$ for the word 苹果.

The first two measures, `PMI` and `Position-specific PMI`, look at the (logged) ratio between observed and expected frequency. Pointwise mutual information (`PMI`) is defined as:

$$\log_2 \left(\frac{\text{observed frequency}}{\text{expected frequency}} \right) \tag{2.2}$$

The `PMI` for the word 苹果 ("apple") therefore is $\log_2 \left(\frac{50.49}{0.20} \right) = 7.97$.

Likewise, Position-specific pointwise mutual information (`Position-specific PMI`) can be calculated using Equation 2.1 and Equation 2.2. However, for position-specific PMI the character frequencies are position-specific. That is, instead of using the overall frequencies of both characters, the frequency of character 1 is defined as the frequency of character 1 in the first position of a two-character word and the frequency of character 2 is defined as the frequency of character 2 in the second position of a two-character word. Given that position-specific character frequency counts are lower than or equal to total character frequency counts, position-specific PMI values are always greater than or equal to standard PMI values.

For example, for the word 苹果 ("apple"), while the frequency of two-character words with character 1 (苹) as the first character (frequency: 50.49) is the same as the overall frequency of character 1 (苹) in two-character words, the frequency of two-character words with character 2 (果) as the second character (frequency: 1,519.01) is somewhat lower than the overall frequency of character 2 (果) in two-character words (frequency: 1,587.82). As a result of this, the value of `Position-specific PMI` for 苹果 (8.03) is higher than the corresponding value of `PMI` (7.97).

As pointed out by Gries (2010, p. 14), "pointwise MI is known to return very high association scores for low-frequency words as well as for technical terms or other expressions that exhibit very little or no variation. On the other hand, the *t*-score returns high association scores to word pairs with high co-occurrence frequencies and provides a better measure of the non-randomness of the co-occurrence" (c.f., Evert, 2009). We therefore included `t-Score` as a third association measure, which is defined as:

$$\frac{\text{observed frequency} - \text{expected frequency}}{\sqrt{\text{expected frequency}}} \tag{2.3}$$

For the word 苹果 ("apple"), this equation results in a *t*-Score of $\frac{50.49-0.20}{\sqrt{0.20}} = 112.06$.

Table 2.7: Pairwise Spearman correlations for the numerical variables in Cluster 2. Abbreviations: PSPMI = Position-specific PMI, TSC = t-Score.

predictor	PMI	PSPMI	TSC
PMI	-		
PSPMI	0.931	-	
TSC	0.910	0.832	-

All three association measures are positive when the observed frequency is greater than the expected frequency and negative when the observed frequency is smaller than the expected frequency. Table 2.7 shows pairwise correlations for the three association measures. All pairwise correlations are strong, positive and significant at the 0.001 α level. This confirms that the three association measures in Cluster 2 encode similar information.

2.4.2.3 Cluster 3: character 1 frequency

As can be seen in Table 2.4, Cluster 3 consists of 12 numerical predictors related to the frequency of the first character. The first 6 measures are 3 frequency and 3 contextual diversity measures from SCCoW, the Gigaword corpus and SUBTLEX-CH: `Character 1 Frequency (SCCoW)`, `Character 1 Frequency (Gigaword)`, `Character 1 Frequency (SUBTLEX-CH)`, `Character 1 CD (SCCoW)`, `Character 1 CD (Gigaword)`, and `Character 1 CD (SUBTLEX-CH)`. These measures are character-level equivalents of the word-level frequency and contextual diversity measures in Cluster 1. As before, frequencies for characters that did not appear in the SCCoW or in the Gigaword corpus were set to *NA*.[3]

The seventh predictor in Cluster 3 is the family size of the first character (`Character 1 Family Size`). This measure is based on the family size measure in English, where family size is defined as the number of morphologically complex words in which a word occurs as a constituent (Schreuder & Baayen, 1997). In word naming in English, words with larger morphological families are named faster than words with smaller morphological families (Baayen et al., 2006; Hendrix, 2016).

[3]We set frequencies for characters that did not appear in the SCCoW or in the Gigaword corpus to *NA* for consistency with the word frequency measures. For word frequency, there was a conceptual reason behind this: the absence of a word in a corpus could either be indicative of true absence of that word or of wrongful segmentation. For character frequency, segmentation issues do not play a role. The absence of a character in a corpus, therefore, is always a true absence. Users of the CLD should therefore feel free to replace *NA* character frequency counts with 0 if they so desire.

For Mandarin Chinese, we define family size of a character as the number of two-character words in the CLD that contain that character.[4] For example, the character 抹 ("to wipe") occurs not only as an independent words, but also is a constituent in 5 two-character words: 抹黑 ("to shame, to smear politically"), 抹布 ("dish towel"), 抹杀 ("to write off"), 抹煞 ("to write off", synonym of 抹杀), and 涂抹 ("to smear, to paint, to scribble"). The family size of the character 抹, therefore, is 5.

A measure based on the family size measure is family frequency (`Character 1 Family Frequency`), which is defined as the summed SCCoW frequency of a character's family members; i.e., the frequency of all two-character words that contain a given character. For the character 抹 ("to wipe") the family frequency is the sum of the frequency of its family members 抹黑 (frequency: 1.85), 抹布 (frequency: 1.44), 抹杀 (frequency: 1.40), 抹煞 (frequency: 0.19), and 涂抹 (frequency: 5.47), which is 10.35.

The cluster of character 1 frequency measures furthermore contains two predictors related to the number of friends of a word: `Character 1 Friends` and `Character 1 Friends Frequency`. The term "friends" refer to the number of words in which the character occurs and is pronounced the same. As mentioned above the character 抹 ("to wipe", "[mä1]") occurs as an independent word, but also is a character in 5 two-character words: 抹黑 ("to shame, to smear politically"), 抹布 ("dish towel"), 抹杀 ('to write off"), 抹煞 ("to write off", synonym of 抹杀), and 涂抹 ("to smear, to paint, to scribble"). In 4 of these words the character is pronounced differently (抹黑, 抹杀, 抹煞, and 涂抹; all "[mɔ3]"), whereas in 1 word it is pronounced the same (抹布; "[mä1]"). Therefore, the number of character 1 friends for the single-character word 抹 is 1. The first characters in the words 抹黑, 抹杀, 抹煞, and 涂抹 each have 3 friends.

`Character 1 Friends Frequency` is the summed frequency of all character 1 friends. For the first character in the word 抹黑 ("to shame, to smear politically"), the summed frequency of the friends 抹杀 (frequency: 1.40), 抹煞 (frequency: 0.19), and 涂抹 (frequency: 5.47) is 7.06. `Character 1 Friends Frequency` for the word 抹黑, therefore, is 7.06.

[4]Note that the definition of a morpheme in Chinese is somewhat problematic. Typically, a character corresponds to a morpheme. Sometimes, however, it can be argued that a morpheme consists of multiple characters (e.g., 刹那 ("very short period of time", "instant")). Furthermore, there is the problem of polysemy: in many cases the meaning of a character depends on the word it appears in (e.g., the character 行 occurs in both 行业 ("business", "profession", first character: "line") and 行为 ("behaviour", first character: "to walk", "to go")). To avoid such issues regarding the nature of a morpheme in Chinese, we use a character-based definition of family size.

The final two numerical variables in Cluster 3 are the number of friends and the frequency of the friends of the phonetic radical (`Character 1 PR Friends`, `Character 1 PR Friends Frequency`). These measures are conceptually similar to the friends and friends frequency measures at the character-level. For a given word, the number of friends of a character's phonetic radical is defined as the number of words that share the phonetic radical and in which the character that contains the phonetic radical is pronounced the same as the character that contains the phonetic radical in the current word.

For example, the phonetic radical for the character 沫 ("foam"; "[mɔ4]") is 末. This phonetic radical occurs in 11 words: 抹黑 ("to shame"), 抹布 ("dish towel"), 茉 ("jasmine"), 茉莉 ("jasmine"), 抹 ("to wipe"), 抹杀 ("to write off"), 抹煞 ("to write off"), 泡沫 ("foam"), 吐沫 ("to spit"), 唾沫 ("saliva"), and 涂抹 ("to smear"). In four of these words, the phonetic radical is pronounced as "[mɔ4]" (茉, 茉莉, 泡沫, and 唾沫). In the other words, it is pronounced as either "[mä1]" (抹布, 抹), "[mɔ3]" (抹黑, 抹杀, 抹煞, 涂抹) or "[mɔ5]" (吐沫). The number of phonological friends of the first (and only) character of the word 沫, therefore, is 4.

The phonetic radical friends frequency is the frequency of the phonetic radical friends. For the word 沫, hence the phonetic radical friends frequency is the summed frequency of the words 茉 (frequency: 0.01), 茉莉 (frequency: 0.81), 泡沫 (frequency: 37.04), and 唾沫 (frequency: 0.59), which is 38.45.

Table 2.8 presents the pairwise correlations between the numerical predictors in Cluster 3. Similar to the first 2 clusters, all the correlations are strong, positive and significant at the 0.001 α level. The lowest pairwise correlation for the measures in Cluster 3 is no less than 0.694 (for the correlation between `Character 1 CD (SUBTLEX-CH)` and `Character 1 PR Friends`). Cluster 3, therefore is a homogeneous cluster.

2.4.2.4 Cluster 4: character 2 frequency

Cluster 4 is the second character counterpart of Cluster 3, which contained 12 frequency measures for the first character. Cluster 4 likewise consists of 12 numerical variables: `Character 2 Frequency (SCCoW)`, `Character 2 Frequency (Gigaword)`, `Character 2 Frequency (SUBTLEX-CH)`, `Character 2 CD (SCCoW)`, `Character 2 CD (Gigaword)`, and `Character 2 CD (SUBTLEX-CH)`, `Character 2 Family Size`, `Character 2 Family Frequency`, `Character 2 Friends`, `Character 2 Friends Frequency`, `Character 2 PR Friends`, and `Character 2 PR Friends Frequency`.

Table 2.8: Pairwise Spearman correlations for the numerical variables in cluster 3. Abbreviations: C1FSC = Character 1 Frequency (SCCoW), C1FGI = Character 1 Frequency (Gigaword), C1FSU = Character 1 Frequency (SUBTLEX-CH), C1CDSC = Character 1 CD (SCCoW), C1CDGI = Character 1 CD (Gigaword), C1CDSU = Character 1 CD (SUBTLEX-CH), C1FS = Character 1 Family Size, C1FF = Character 1 Family Frequency, C1FR = Character 1 Friends, C1FRF = Character 1 Friends Frequency, C1PFR = Character 1 PR Friends, C1PFRF = Character 1 PR Friends Frequency.

predictor	C1FSC	C1FGI	C1FSU	C1CDSC	C1CDGI	C1CDSU	C1FS	C1FF	C1FR	C1FRF	C1PFR	C1PFRF
C1FSC	-											
C1FGI	0.987	-										
C1FSU	0.884	0.879	-									
C1CDSC	0.990	0.977	0.892	-								
C1CDGI	0.984	0.991	0.887	0.990	-							
C1CDSU	0.888	0.875	0.990	0.905	0.893	-						
C1FS	0.856	0.847	0.838	0.849	0.839	0.839	-					
C1FF	0.969	0.949	0.847	0.964	0.951	0.858	0.858	-				
C1FR	0.835	0.827	0.815	0.827	0.818	0.815	0.979	0.839	-			
C1FRF	0.966	0.948	0.853	0.961	0.950	0.862	0.846	0.966	0.859	-		
C1PFR	0.727	0.720	0.694	0.719	0.713	0.694	0.834	0.726	0.854	0.746	-	
C1PFRF	0.856	0.841	0.730	0.849	0.841	0.735	0.738	0.858	0.752	0.882	0.845	-

The pairwise correlations for the measures in Cluster 4 are presented in Table 2.9. As was the case for the correlations in Cluster 3, all correlations are strong, positive and significant at the 0.001 α level. The weakest correlation is 0.676 (between `Character 2 Frequency (SUBTLEX-CH)` and `Character 2 PR Friends`). Both clusters 3 and 4, therefore, are highly homogeneous clusters. In addition, the pairwise correlations for the numerical variables Cluster 3 are highly similar to the pairwise correlations for the numerical variables in Cluster 4 ($r = 0.948$). The distributional structure of the character 2 frequency measures thus is comparable to that of the character 1 frequency measures.

2.4.2.5 Cluster 5: character 1 entropy

As can be seen in Table 2.4, Cluster 5 consists of 2 numerical predictors: `Character 1 Entropy` and `Character 1 Trigram Entropy`. `Character 1 Entropy` is the entropy over the probability distribution of two-character words, in which the first character is the first character in the current word. It is a measure of the uncertainty about the second character given the first character of a word.

The first character of the word 挤压 ("to squeeze and press", frequency: 8.84), for instance, is 挤. This character is the first character in two other 2-character words: 挤兑 ("panicky bank withdrawal", frequency: 0.97) and 挤占 ("to occupy as a crowd", frequency: 3.75). Converting the frequency counts for these words to probabilities results in a probability of 0.65 for 挤压, a probability of 0.07 for 挤兑 and a probability of 0.28 for 挤占. The entropy ($-\sum_{i=1}^{n} p_i * \log_2(p_i)$) over this probability distribution is 1.19. `Character 1 Entropy` for the word 挤压, therefore, is 1.19.

`Character 1 Trigram Entropy` is conceptually similar to `Character 1 Entropy` and is defined as the entropy over the probability distribution of all character trigrams in the SCCoW in which the first character of the current word is the middle character. Both `Character 1 Entropy` and `Character 1 Trigram Entropy` gauge the combinatorial properties of a character with other characters. `Character 1 Entropy` taps into these combinatorial properties at the word level, whereas `Character 1 Trigram Entropy` looks at combinatorial properties both at and above the word level. Accordingly, the pairwise correlation between `Character 1 Entropy` and `Character 1 Trigram Entropy` is positive and highly significant (0.674, $p < 0.001$).

Table 2.9: Pairwise Spearman correlations for the numerical variables in Cluster 4. Abbreviations: C2FSC = Character 2 Frequency (SCCoW), C2FGI = Character 2 Frequency (Gigaword), C2FSU = Character 2 Frequency (SUBTLEX-CH), C2CDSC = Character 2 CD (SCCoW), C2CDGI = Character 2 CD (Gigaword), C2CDSU = Character 2 CD (SUBTLEX-CH), C2FS = Character 2 Family Size, C2FF = Character 2 Family Frequency, C2FR = Character 2 Friends, C2FRF = Character 2 Friends Frequency, C2PFR = Character 2 PR Friends, C2PRFRF = Character 2 PR Friends Frequency.

predictor	C2FSC	C2FGI	C2FSU	C2CDSC	C2CDGI	C2CDSU	C2FS	C2FF	C2FR	C2FRF	C2PFR	C2PRFRF
C2FSC	-											
C2FGI	0.981	-										
C2FSU	0.835	0.832	-									
C2CDSC	0.985	0.966	0.846	-								
C2CDGI	0.975	0.984	0.840	0.987	-							
C2CDSU	0.844	0.829	0.987	0.869	0.853	-						
C2FS	0.803	0.793	0.797	0.792	0.779	0.798	-					
C2FF	0.971	0.946	0.796	0.960	0.944	0.812	0.799	-				
C2FR	0.770	0.761	0.756	0.757	0.745	0.756	0.963	0.770	-			
C2FRF	0.953	0.931	0.786	0.943	0.929	0.800	0.774	0.952	0.803	-		
C2PFR	0.717	0.706	0.676	0.702	0.694	0.677	0.861	0.709	0.892	0.745	-	
C2PFRF	0.885	0.862	0.712	0.871	0.859	0.723	0.722	0.882	0.748	0.923	0.809	-

2.4.2.6 Cluster 6: character 2 entropy

Cluster 6 is the character 2 counterpart of cluster 5 and contains the predictors `Character 2 Entropy` and `Character 2 Trigram Entropy`. In addition, it contains a third predictor: `Character 2 PR Frequency`. `Character 2 PR Frequency` is the frequency of the phonetic radical of the second character. The frequency of the phonetic radical is defined as the summed frequency of all characters that contain that phonetic radical. The phonetic radical for the character 沫, for instance, is 末. This phonetic radical occurs in 3 characters: 抹 (frequency: 14.77), 茉 (frequency: 1.51), and 沫 (frequency: 27.95). The frequency of the phonetic radical of the second character of the word 唾沫 ("saliva") is the sum of the frequencies of these 3 characters, which is 44.23.

The frequency of the phonetic radical of the first character clusters with measures of the homography of the phonetic radical of the first character (see Cluster 17). By contrast, the frequency of the phonetic radical of the second character clusters with the entropy and trigram entropy of the second character. This is somewhat surprising, given the fact that `Character 2 PR Frequency` correlates more strongly with measures of the homography of the phonetic radical of the second character (see Cluster 18; `Character 2 PR Enemies Frequency`: $r = 0.721$, `Character 2 PR Enemies (tokens)`: $r = 0.651$, `Character 2 PR Enemies (types)`: $r = 0.536$) than with `Character 2 Entropy` ($r = 0.389$) and `Character 2 Trigram Entropy` ($r = 0.400$). Indeed, a number of alternative SOMs with the same structure, but a different random initialization of weights showed a clustering of `Character 2 PR Frequency` with `Character 2 PR Enemies Frequency`, `Character 2 PR Enemies (Tokens)` and `Character 2 PR Enemies (Tokens)`. To some extent the organization of the SOM and the outcome of the clustering algorithm thus depend on the random initialization of weights. Nonetheless, all pairwise correlations for the measures in Cluster 6 (see Table 2.10) are positive and significant at an α level of 0.001. Despite the above considerations, therefore, the 3 measures in Cluster 6 form a reasonably homogeneous cluster.

2.4.3 Group 2: visual complexity measures

Group 2 consists of two clusters, Cluster 7 and Cluster 8, which are shown in red in Figure 2.3. Both of these clusters contain visual complexity measures. The numerical variables in Cluster 7 describe the visual complexity of character 1,

Table 2.10: Pairwise Spearman correlations for the numerical variables in Cluster 6. Abbreviations: C2H = Character 2 Entropy, C2H3 = Character 2 Trigram Entropy, C2PRF = Character 2 PR Frequency.

predictor	C2H	C2H3	C2PRF
C2H	-		
C2H3	0.643	-	
C2PRF	0.389	0.400	-

whereas the numerical variables in Cluster 8 encode information about the visual complexity of character 2. An overview of the numerical variables in Cluster 7 and Cluster 8 can be seen in Table 2.11.

2.4.3.1 Cluster 7: character 1 visual complexity

As can be seen in Table 2.11, Cluster 7 consists of 9 measures: `Character 1 Strokes`, `Character 1 High-Level Components`, `Character 1 Low-Level Components`, `Character 1 Pixels`, `Character 1 Picture Size`, `Character 1 Low-Level Components N`, `Character 1 Low-Level Components OLD`, `Character 1 Pixels OLD`, and `Character 1 PR Strokes`.

`Character 1 Strokes` is the number of strokes of the first character. The number of strokes in a character is a common measure of the complexity of Chinese characters. A stroke refers to a line that is written continuously without a pause. As noted by Zheng (1983) (see also Perfetti & Tan, 1999), 24 different strokes exist in the Chinese writing system. The greater the number of strokes a character consists of, the greater the visual complexity of that character.

Visual complexity can be described at multiple grain sizes. In addition to number of strokes, the CLD contains two measures that describe the number of visual components in a word at a *larger* grain size: `Character 1 High-Level Components` and `Character 1 Low-Level Components`. Both of these measures were developed in the context of exploring the potential of the naive discrimination learning framework (see Baayen et al., 2011) for Chinese and describe recurring visual patterns that are spatially separable.

High-level components describe visual complexity at a relatively large grain size. The high-level visual components of first character of the word 欣喜 ("happy"), for instance, are 斤 and 欠 (where 斤 is on the left side and 欠 on the right side of the character). For the low-level visual components measures, the high-level

Table 2.11: Overview of numerical predictors: visual complexity (Group 2)

	mean	median	sd	min	max	NA
Cluster 7: character 1 visual complexity						
C1 Strokes	8.66	8	3.48	1	27	0
C1 High-Level Comp.	2.71	3	1.14	1	10	3
C1 Low-Level Comp.	5.80	6	2.31	1	18	3
C1 Pixels	3598.04	3686	621.99	957	5193	0
C1 Picture Size	2933.65	3001	830.16	649	5761	0
C1 LLC N	1.70	0	3.64	0	27	3
C1 LLC OLD	2.52	2.40	0.96	0.90	10.95	3
C1 Pixels OLD	2506.20	2497.20	285.46	1520.90	3588.60	0
C1 PR Strokes	6.57	6	2.74	1	22	14478
Cluster 8: character 2 visual complexity						
C2 Strokes	8.34	8	3.37	1	25	4710
C2 High-Level Comp.	2.61	2	1.11	1	9	4711
C2 Low-Level Comp.	5.60	6	2.20	1	17	4711
C2 Pixels	3541.54	3629	625.89	957	5185	4710
C2 Picture Size	2891.27	2952	832.18	649	5464	4710
C2 LLC N	1.57	0	3.28	0	27	4711
C2 LLC OLD	2.46	2.35	0.89	0.90	8.00	4711
C2 Pixels OLD	2491.64	2497.30	286.88	1520.90	3529.10	4710
C2 PR Strokes	6.33	6	2.65	1	22	17944
Strokes	15.71	16	5.19	1	42	0

visual components are further decomposed into visually separable patterns at or just above the stroke level. The character 欣, for instance, is decomposed into 4 low-level components. The left high-level component of the character 欣 (斤) is decomposed into the low-level components 厂 and 𠂇, whereas the right part of the character 欣 (欠) is decomposed into 𠂉 and 人. The measures `Character 1 High-Level Components` and `Character 1 Low-Level Components` refer to the number of high-level components and low-level components, respectively.

Visual complexity can also be described at a *smaller* grain size than strokes. The CLD provides two measures of visual complexity at such a small grain size: `Character 1 Pixels` and `Character 1 Picture Size`. To calculate `Character 1 Pixels`, we generated a PNG image file with the character in black (font: SimHei,

font size: 80) centered on a 150 by 150 pixel white background for each of the 5,242 unique characters in the CLD. `Character 1 Pixels` is the number of non-white pixels in the image file for the first character in a word. For example, the image file for the first character 财 in the word 财税 ("taxation") contains 3,801 non-white pixels (out of the total $150^2 = 22,500$ pixels). `Character 1 Pixels` for the word 财税, hence is 3,801. `Character 1 Picture Size` likewise taps into the visual complexity of the PNG files for each character and is defined as the size of the image file in bytes. For example, the image file for the first character 财 in the word 财税 is 2,960 bytes. `Character 1 Picture Size` for the word 财税, therefore, is 2,960.

Cluster 7 furthermore contains 3 numerical predictors related to neighbourhood characteristics of the visual features of a character. The first two of these measures, `Character 1 Low-Level Components N` and `Character 1 Low-Level Components OLD`, are at the level of the low-level components described above. `Character 1 Low-Level Components N` is the number of characters at a Hamming distance of 1 (i.e., 1 different low-level component; same number of low-level components) from the first character. `Character 1 Low-Level Components OLD` is the average orthographic Levenshtein distance (OLD, i.e., the number of deletions, insertions, or substitutions necessary to get from the low-level components of the target character to the low-level component of another character) of the 20 closest neighbours of the first character.

The third measure that describes neighbourhood characteristics of the visual features of a character is `Character 1 Pixels OLD`: the orthographic Levenshtein distance of the 20 closest neighbours at the pixel level. As mentioned above, each character consists of black pixels on a 150 by 150 pixels white background. Prior to calculating the OLD between characters, we set all white pixels to 0 and all non-white pixels to 1. This resulted in a 150 by 150 matrix of zeroes and ones for each character. We then defined the OLD between two characters as the summed difference between the matrices for both characters. For each character, `Character 1 Pixels OLD` is the average of the OLD between that character and the 20 closest neighbours.

For the word 一生 ("lifetime"), for instance, the 20 characters with the smallest Levenshtein distance to the first character 一, are 千 (distance: 1,444), 干 (distance: 1,437), 十 (distance: 1,766), 卜 (distance: 1,780), 卞 (distance: 1,832), 于 (distance: 1,873), 子 (distance: 1,908), 宁 (distance: 2,017), 壬 (distance: 2,029), 三 (distance: 2,057), 予 (distance: 2,072), 产 (distance: 2,083), 丫 (distance:

2,180), 守 (distance: 2,211), 心 (distance: 2,222), 开 (distance: 2,225), 分 (distance: 2,226), 兴 (distance: 2,235), 七 (distance: 2,237), and 广 (distance: 2,253). `Character 1 Pixels OLD` for the word 一生 is the average of these 20 distances, which is 2,004.35.

The final measure of the visual complexity of the first character is the number of strokes of the phonetic radical: `Character 1 PR Strokes`. For example, the first character of the word 但是 ("but") is 但. The phonetic radical of this character is 旦. This phonetic radical consists of 5 strokes. `Character 1 PR Strokes` for the word 但是, therefore, is 5.

Table 2.12 shows the pairwise correlations for the predictors in cluster 7. For the clusters in Group 1 (frequency measures), all pairwise correlations were positive. Here, however, we see some negative correlations as well. To be precise, `Character 1 Low-Level Components N` is negatively correlated with all other measures in cluster 7. This is desirable behaviour of the clustering algorithm, because both positive and negative correlations between predictors indicate that the values of one predictor are not independent of the values of the other predictor. The fact that predictors with both positive and negative correlations can cluster together is an advantage of using the *squared* correlation matrix as input data to the SOM described above, rather than the correlation matrix itself. As before, all pairwise correlations for the measures in Cluster 7 are significant at the 0.001 α level.

2.4.3.2 Cluster 8: character 2 visual complexity

The bottom half of Table 2.11 provides an overview of the numerical variables in Cluster 8. In total, Cluster 8 consists of 10 measures. The first 9 are the character 2 counterparts of the numerical variables in cluster 7: `Character 2 Strokes`, `Character 2 High-Level Components`, `Character 2 Low-Level Components`, `Character 2 Pixels`, `Character 2 Picture Size`, `Character 2 Low-Level Components N`, `Character 2 Low-Level Components OLD`, `Character 2 Pixels OLD`, and `Character 2 PR Strokes`. The last predictor `Strokes` is the number of strokes in the word as a whole (i.e., the sum of the number of strokes in the first and second character jointly). The fact that the number of strokes of the word as a whole clusters with the number of strokes of the second character rather than with the number of strokes of the first character is unsurprising, given the fact that the (Spearman) correlation between `Strokes` and `Character 2 Strokes` ($r = 0.700$) is stronger than the correlation between `Strokes` and `Character 1 Strokes` ($r = 0.540$).

Table 2.12: Pairwise Spearman correlations for the numerical variables in Cluster 7. Abbreviations: C1S = Character 1 Strokes, C1HC = Character 1 High-Level Components, C1LC = Character 1 Low-Level Components, C1P = Character 1 Pixels, C1PS = Character 1 Picture Size, C1LCN = Character 1 Low-Level Components N, C1LCOLD = Character 1 Low-Level Components OLD, C1POLD = Character 1 Pixels OLD, C1PRS = Character 1 PR Strokes.

predictor	C1S	C1HC	C1LC	C1P	C1PS	C1LCN	C1LCOLD	C1POLD	C1PRS
C1S	-								
C1HC	0.786	-							
C1LC	0.824	0.750	-						
C1P	0.841	0.698	0.772	-					
C1PS	0.483	0.494	0.573	0.466	-				
C1LCN	−0.534	−0.521	−0.659	−0.510	−0.537	-			
C1LCOLD	0.723	0.682	0.875	0.673	0.674	−0.791	-		
C1POLD	0.592	0.527	0.549	0.586	0.540	−0.481	0.619	-	
C1PRS	0.799	0.667	0.639	0.648	0.279	−0.280	0.504	0.422	-

Table 2.13: Pairwise Spearman correlations for the numerical variables in Cluster 8. Abbreviations: C2S = Character 2 Strokes, C2HC = Character 2 High-Level Components, C2LC = Character 2 Low-Level Components, C2P = Character 2 Pixels, C2PS = Character 2 Picture Size, C2LCN = Character 2 Low-Level Components N, C2LCOLD = Character 2 Low-Level Components OLD, C2POLD = Character 2 Pixels OLD, C2PRS = Character 2 PR Strokes, S = Strokes.

predictor	C2S	C2HC	C2LC	C2P	C2PS	C2LCN	C2LCOLD	C2POLD	C2PRS	S
C2S	-									
C2HC	0.784	-								
C2LC	0.818	0.751	-							
C2P	0.836	0.690	0.772	-						
C2PS	0.462	0.483	0.564	0.452	-					
C2LCN	−0.522	−0.488	−0.637	−0.489	−0.496	-				
C2LCOLD	0.715	0.667	0.861	0.667	0.644	−0.786	-			
C2POLD	0.584	0.517	0.537	0.573	0.526	−0.449	0.603	-		
C2PRS	0.789	0.668	0.625	0.636	0.247	−0.287	0.501	0.434	-	
S	0.700	0.555	0.578	0.583	0.337	−0.371	0.513	0.420	0.524	-

Table 2.13 shows the pairwise correlations for the variables in Cluster 8. Similar to `Character 1 Low-Level Components N` in cluster 7, `Character 2 Low-Level Components N` is negatively correlated with the other variables in Cluster 8. All pairwise correlations are significant at the 0.001 α level, including the pairwise correlations between the strokes of the word as a whole and all other variables. The pairwise correlations between the measures in Cluster 7 are nearly perfectly correlated with the pairwise correlations for the character 2 measures in Cluster 8 ($r > 0.999$). The distributional space for the visual complexity of a character is therefore almost identical for characters 1 and 2.

2.4.4 Group 3: phonological measures

Group 3 consists of 6 clusters that describe phonological properties of the words in the CLD (see Table 2.14). Phonological frequencies in Table 2.14 were rounded to 1 decimal place to prevent the table from exceeding the page width. The numerical variables in Cluster 9 and Cluster 10 describe the phonological complexity of character 1 and character 2 and furthermore contain measures about the frequency of the diphones in a word and its characters. Clusters 11 and 12 describe phonological neighbourhood characteristics. Finally, the measures in Clusters 13 and 14 describe the phonological frequency of both characters, as well as the phonological frequency of the word as a whole. As can be seen in Figure 2.3, the 6 clusters in Group 3 occupy a spatially continuous area on the left side of the SOM, with character 1 measures more towards the top of the map and character 2 measures more towards the bottom of the map.

Table 2.14: Overview of numerical predictors: phonological measures (Group 3)

	mean	median	sd	min	max	NA
Cluster 9: character 1 phonological complexity						
C1 Phonemes	2.79	3	0.70	1	4	0
C1 Mean Diph. Freq.	31221.7	25966.4	20284.5	48.7	110568.3	729
C1 Max Diph. Freq.	46254.0	39120.0	32502.0	48.7	110568.3	729
Mean Diphone Freq.	27194.0	24349.9	14123.1	48.7	110568.3	134
Max Diphone Freq.	60086.0	55036.7	32759.4	48.7	110568.3	134

Table 2.14 (continued)

	mean	median	sd	min	max	NA
Cluster 10: character 2 phonological complexity						
C2 Phonemes	2.76	3	0.71	1	4	4710
C2 Mean Diph. Freq.	31469.8	25990.6	20365.0	48.7	110568.3	5307
C2 Max Diph. Freq.	45661.7	39120.0	32298.2	48.7	110568.3	5307
Cluster 11: character 1 phonological neighbourhood						
C1 Phonological N	15.54	16	5.58	1	28	0
C1 PLD	1.17	1.10	0.19	1.00	1.90	0
Cluster 12: character 2 phonological neighbourhood						
C2 Phonological N	15.68	16	5.67	1	28	4710
C2 PLD	1.17	1.10	0.19	1.00	1.90	4710
Phonological N	3.82	2	5.56	0	27	0
PLD	1.99	1.95	0.49	1.00	3.65	0
Phonemes	5.12	5	1.40	1	8	0
Cluster 13: character 1 phonological frequency						
C1 Mean Phon. Freq.	155039.7	147559.2	60450.6	25913.7	370975.6	0
C1 Min Phon. Freq.	64183.0	48523.4	54441.7	3304.0	370975.6	0
C1 Max Phon. Freq.	257868.5	231777.4	101007.3	48523.4	370975.6	0
C1 Init. Phon. Freq.	85954.1	57216.5	92211.1	12595.9	370975.6	0
C1 Min Diph. Freq.	17023.6	11400.6	17469.1	0.0	110568.3	729
C1 Init. Diph. Freq.	19894.6	11731.4	22140.9	0.0	110568.3	729
Cluster 14: character 2 phonological frequency						
C2 Mean Phon. Freq.	159663.2	153167.3	60931.1	23184.5	370975.6	4710
C2 Min Phon. Freq.	64820.2	48523.4	55049.4	3304.0	370975.6	4710
C2 Max Phon. Freq.	266566.7	231777.4	100401.2	23184.5	370975.6	4710
C2 Init. Phon. Freq.	87041.9	57216.5	93648.4	12595.9	370975.6	4710
C2 Min Diph. Freq.	18165.2	11731.4	17879.2	5.5	110568.3	5307
C2 Init. Diph. Freq.	20951.6	11799.6	22209.1	5.5	110568.3	5307
Mean Phoneme Freq.	155159.4	153984.3	44727.3	25913.7	370975.6	0
Min Phoneme Freq.	44455.4	36567.9	32392.2	3304.0	370975.6	0
Max Phoneme Freq.	307096.7	370975.6	85554.6	48523.4	370975.6	0
Min Diphone Freq.	4974.9	2332.4	8929.9	0.0	110568.3	134
Trans. Diph. Freq.	3753.0	2659.1	5349.2	0.1	110568.3	4710

2.4.4.1 Cluster 9: character 1 phonological complexity

Cluster 9 consists of 5 measures: one for the phonological complexity of the first character (`Character 1 Phonemes`), two for the diphone frequency of the first character (`Character 1 Mean Diphone Frequency`, `Character 1 Max Diphone Frequency`) and two for the diphone frequency of the word as a whole (`Mean Diphone Frequency`, `Max Diphone Frequency`). Although Cluster 9 contains a single measure of phonological complexity only, we refer to this cluster as describing "character 1 phonological complexity" to avoid confusion with Cluster 13, which consists solely of phonological frequency measures for the first character.

`Character 1 Phonemes` is the phoneme count of the first character. This count is based on the IPA transcriptions introduced in Section 2.3. In Chinese, a relatively large number of "double vowel" sequences occur. These diphthongs can be considered as either single phonemes or as phoneme sequences, potentially as a function of the pitch contour (rising or falling). Hayes (2009) argues that considering diphthongs in Mandarin Chinese as phoneme sequences has the advantage of needing a smaller inventory of phonemes and fitting well with the phonological process of assimilation. We therefore opted to consider diphthongs as a sequence of two phonemes for the phoneme counts reported here.

Cluster 9 furthermore contains 4 measures of the frequency of the diphones in the first character of a word, as well as in the word as a whole. For the regularity of the phonetic radical (`Character 1 PR Regularity`, `Character 2 PR Regularity`) and measures of the phonological neighbourhood density (see below) we take tonal differences into account. Tone is a phonological property at the suprasegmental level: it contains information about the phonological similarity of the pronunciation of different characters above the level of segments. Tones therefore provide important information for the overall similarity of character pronunciations. Phoneme and diphone frequency measures, however, describe phonological properties of the character at the segmental level. For phoneme and diphone frequency measures, we therefore ignore tonal differences.

The frequency of a diphone is defined as the summed frequency of all words in the CLD in which a phoneme occurs. `Character 1 Mean Diphone Frequency` is the average diphone frequency of the first character in a word. The pronunciation "[meɪ2]" of the first character 玫 of the word 玫瑰 ("rose"), for instance, consists of 2 diphones: "[me]" (frequency: 4026.14), and "[eɪ]"(frequency: 26533.89). For the

word 玫瑰, therefore, `Character 1 Mean Diphone Frequency` is $\frac{4{,}026.14+26{,}533.89}{2} = 15,280.02$. `Character 1 Max Diphone Frequency` is the maximum frequency of the diphones in the first character. The most frequent diphone in the first character 玫 of the word 玫瑰 is [ei] (frequency: 26533.89). `Character 1 Max Diphone Frequency` for the word 玫瑰, thus, is 26533.89.

From a conceptual perspective, one might expect the average and the maximum diphone frequency of the first character to cluster with the minimum and initial diphone frequency of the first character, rather than with `Character 1 Phonemes`. The correlation matrix for the CLD, however, suggests that the current clustering makes a lot of sense. The 4 predictors that show the highest correlation with `Character 1 Mean Diphone Frequency` are `Character 1 Max Diphone Frequency` ($r = 0.931$), `Mean Diphone Frequency` ($r = 0.724$), `Max Diphone Frequency` ($r = 0.627$), and `Character 1 Phonemes` ($r = 0.551$). Similarly, the 4 predictors that have the highest correlation with `Character 1 Max Diphone Frequency` are `Character 1 Mean Diphone Frequency` ($r = 0.931$), `Mean Diphone Frequency` ($r = 0.688$), `Character 1 Phonemes` ($r = 0.683$), and `Max Diphone Frequency` ($r = 0.675$). The data in the CLD, therefore, do not support the intuition that all diphone frequency measures for the first character should cluster together.

`Mean Diphone Frequency` is the whole-word counterpart of `Character 1 Mean Diphone Frequency` and is defined as the average frequency of the diphones in the word as a whole. Analogously, `Max Diphone Frequency` is the frequency of the most frequent diphone in the word as a whole. The mean and maximum frequency of the diphones in a word as a whole are clustered with the corresponding first character measures, rather than with the corresponding second character measures. This is unsurprising given the presence of single character words in the CLD and the correlations of `Mean Diphone Frequency` (`Character 1 Mean Diphone Frequency`: $r = 0.724$, `Character 2 Mean Diphone Frequency`: $r = 0.664$) and `Mean Diphone Frequency` (`Character 1 Max Diphone Frequency`: $r = 0.675$, `Character 2 Max Diphone Frequency`: $r = 0.609$) with the corresponding character 1 and character 2 measures. Again, therefore, the clustering algorithm correctly follows the correlational structure of the data in the CLD.

Table 2.15 shows the pairwise correlations for the measures in Cluster 9. As before, all correlations are positive and highly significant at the 0.001 α level. The minimum correlation (i.e., the correlation between `Character 1 Phonemes` and `Mean Diphone Frequency`) is no less than 0.430. This confirms that the numerical variables in Cluster 9 form a homogeneous cluster.

Table 2.15: Pairwise Spearman correlations for the numerical variables in Cluster 9. Abbreviations: C1P = Character 1 Phonemes, C1MEDF = Character 1 Mean Diphone Frequency, C1MAXDF = Character 1 Max Diphone Frequency, MEDF = Mean Diphone Frequency, MAXDF = Max Diphone Frequency.

predictor	C1P	C1MEDF	C1MAXDF	MEDF	MAXDF
C1P	-				
C1MEDF	0.551	-			
C1MAXDF	0.683	0.931	-		
MEDF	0.430	0.724	0.688	-	
MAXDF	0.439	0.627	0.675	0.871	

2.4.4.2 Cluster 10: character 2 phonological complexity

Cluster 10 is the character 2 counterpart of Cluster 9 and consists of 3 measures of the phonological complexity (`Character 2 Phonemes`) and the diphone frequency (`Character 2 Mean Diphone Frequency`, `Character 2 Max Diphone Frequency`) of the second character. As before, despite the fact that Cluster 10 contains a single measure of phonological complexity only, we refer to this cluster as describing "character 2 phonological complexity" to avoid confusion with Cluster 14, which contains phonological frequency measures for the second character.

As was the case for Cluster 9, the numerical variables in Cluster 10 form a homogeneous cluster. All pairwise correlations are positive and highly significant at an α level of 0.001 (see Table 2.16). Furthermore, the pairwise correlations for the measures of the second character in Cluster 10 show a near-perfect correlation with the pairwise correlations for the first character measures in Cluster 9 ($r > 0.999$). This demonstrates that the distributional structure for the phonological complexity measures of the first and second character is highly similar.

Table 2.16: Pairwise Spearman correlations for the numerical variables in Cluster 10. Abbreviations: C2P = Character 2 Phonemes, C2MEDF = Character 2 Mean Diphone Frequency, C2MAXDF = Character 2 Max Diphone Frequency.

predictor	C2P	C2MEDF	C2MAXDF
C2P	-		
C2MEDF	0.540	-	
C2MAXDF	0.687	0.932	

2.4.4.3 Cluster 11: character 1 phonological neighbourhood

Cluster 11 consists of two measures of the phonological neighbourhood of the first character: `Character 1 Phonological N` and `Character 1 PLD`. `Character 1 Phonological N` describes the number of phonological neighbours for the first character of a word. A phonological neighbour is defined as a pronunciation that differs by one phoneme from the pronunciation of the target word (i.e., the Hamming distance between a pronunciation and the target pronunciation is 1). We use a strict definition of neighbourhood, in the sense that tone is considered when determining the Hamming distance between two pronunciations. The pronunciation of the first character 中 in the word 中东 ("Middle East"), for instance, is [ʈʂɔŋ1]. This pronunciation has 15 phonological neighbours: [tʰɔŋ1], [tʰsɔŋ1], [tɔŋ1], [kɔŋ1], [kʰɔŋ1], [sɔŋ1], [xɔŋ1], [lɔŋ1], [iɔŋ1], [tsɔŋ1], [ʈʂɑŋ1], [ʈʂʏŋ1], [ʈʂɔŋ3], and [ʈʂɔŋ4]. For the word 中东, `Character 1 Phonological N`, therefore, is 15.

`Character 1 PLD` is the average phonological Levenshtein distance of the 20 closest phonological neighbours for the first character. For the pronunciation of the first character 中 in the word 中东, for instance, the 20 pronunciations with the smallest edit distance include the 15 phonological neighbours mentioned above (edit distance 1), as well as 5 of the 100 character pronunciations at an edit distance of 2 (e.g., [mɔ1], [ɔ1], [sɑŋ1], [iɔŋ3], [nɔŋ2], ...). `Character 1 PLD` for the word "中东", therefore, is $\frac{15*1+5*2}{20} = 1.25$.

The correlation between `Character 1 Phonological N` and `Character 1 PLD`, is negative ($r = -0.826$). The greater the number of phonological neighbours at a Hamming distance of 1, the smaller the average distance between the target character and its 20 closest phonological neighbours. Again, this highlights the advantage of using the *squared* correlation matrix as input data to the SOM (as opposed to the raw correlation matrix), which allows negatively correlated predictors to have a similar topographical distribution across the SOM.

2.4.4.4 Cluster 12: character 2 phonological neighbourhood

Cluster 12 consists of 5 numerical variables. The first two measures, `Character 2 Phonological N` and `Character 2 PLD` are the character 2 counterparts of the phonological neighbourhood measures for the first character in Cluster 11. `Phonological N` and `PLD` are analogous measures for the word as a whole.

Measures of the average and maximum diphone frequency of the word as a whole clustered with the corresponding measures for the first character. By contrast, phonological neighbourhood measures for the word cluster with phonological neighbourhood measures for the second character. Again, the clustering algorithm and the underlying SOM follow the distributional properties of the data: for both `Phonological N` (`Character 1 Phonological N`: $r = 0.182$, `Character 2 Phonological N`: $r = 0.243$) and `PLD` (`Character 1 PLD`: $r = 0.239$, `Character 2 PLD`: $r = 0.334$) the correlation of the word-level measure with the corresponding second character measure is greater than the correlation of the word-level measure with the corresponding first character measure.

The final measure in Cluster 12 is `Phonemes`, which is the sum of `Character 1 Phonemes` and `Character 2 Phonemes`. As can be seen in Table 2.17, `Phonemes` shows medium-strength correlations with the other measures in Cluster 12. The strongest pairwise correlation for `Phonemes`, however, is with `Character 2 Phonemes` ($r = 0.720$). Furthermore, the average absolute value for the correlation of `Phonemes` with the other measures in Cluster 12 is 0.244, whereas the average absolute value of the correlation of `Phonemes` with the numerical variables in Cluster 10 is 0.303. Based on these observations, one might expect `Phonemes` to be a part of Cluster 10, rather than Cluster 12. Indeed, in a number of similar SOMs with a different random initialization of weights `Phonemes` was clustered with the measures in Cluster 10. Nonetheless, all pairwise correlations for the measures in Cluster 12 are significant at an α level of 0.001 (see Table 2.17). Although the composition of the cluster may not be optimal, therefore, the measures in Cluster 12 form a fairly homogeneous cluster.

Table 2.17: Pairwise Spearman correlations for the numerical variables in Cluster 12. Abbreviations: C2PN = Character 2 Phonological N, C2PLD = Character 2 PLD, PN = Phonological N, P = Phonemes.

predictor	C2PN	C2PLD	PN	PLD	P
C2PN	-				
C2PLD	−0.834	-			
PN	0.243	−0.226	-		
PLD	−0.274	0.334	−0.864	-	
P	−0.283	0.374	−0.561	0.664	

2.4.4.5 Cluster 13: character 1 phonological frequency

The 6 measures in Cluster 13 describe the phonological frequency of the first character of a word. The first 4 measures refer to the frequency of the phonemes in this character: `Character 1 Mean Phoneme Frequency`, `Character 1 Min Phoneme Frequency`, `Character 1 Max Phoneme Frequency`, and `Character 1 Initial Phoneme Frequency`. As mentioned in the discussion of Cluster 9, the first character 玫 of the word 玫瑰 ("rose") is pronounced as "[meɪ2]". The pronunciation "[meɪ2]" consists of 3 phonemes: "[m]", "[e]", and "[ɪ]". The frequency of a phoneme is defined as the summed frequency of all words in the CLD in which a phoneme occurs. The frequencies of the phonemes [m], [e], and [ɪ] are 29,457.15, 26,533.89 and 65,558.42, respectively. For the word 玫瑰, therefore, `Character 1 Mean Phoneme Frequency` is $\frac{29{,}457.15+26{,}533.89+65{,}558.42}{3} = 40{,}516.48$, `Character 1 Min Phoneme Frequency` is 26,533.89, `Character 1 Max Phoneme Frequency` is 65,558.42 and `Character 1 Initial Phoneme Frequency` is 29,457.15.

The last 2 measures in Cluster 13 are the initial diphone frequency and the minimum diphone frequency. As noted above in the discussion of Cluster 9, the character 玫 ("[meɪ2]") consists of 2 diphones: "[me]" (frequency: 4026.14), and "[eɪ]" (frequency: 26533.89). Thus, for the word 玫瑰 both `Character 1 Min Diphone Frequency` and `Character 1 Initial Diphone Frequency` are 4,026.14.

Unlike the average and maximum diphone frequency, the minimum and initial diphone frequency of the first character cluster with phoneme frequency measures. As before, the outcome of the clustering algorithm is in line with the distributional structure of the data. The 4 predictors that show the highest correlation with `Character 1 Min Diphone Frequency` are `Character 1 Initial Diphone Frequency` ($r = 0.975$), `Character 1 Mean Phoneme Frequency` ($r = 0.700$), `Character 1 Max Phoneme Frequency` ($r = 0.663$), and `Character 1 Min Phoneme Frequency` ($r = 0.581$). The 4 predictors that have the highest correlation with `Character 1 Initial Diphone Frequency` are `Character 1 Min Diphone Frequency` ($r = 0.975$), `Character 1 Mean Phoneme Frequency` ($r = 0.699$), `Character 1 Max Phoneme Frequency` ($r = 0.676$), and `Character 1 Min Phoneme Frequency` ($r = 0.573$).

Table 2.18 presents the pairwise correlations of variables in Cluster 13. All correlations are positive and significant at the 0.001 α level. The strong correlation between the initial and minimum diphone frequency ($r = 0.975$) is a consequence of the fact that for 26,157 of the 30,645 words in the CLD (85.35%) the initial diphone of the first character is the least frequent diphone in that character. Similarly, the

Table 2.18: Pairwise Spearman correlations for the numerical variables in Cluster 13. Abbreviations: C1MEPF = Character 1 Mean Phoneme Frequency, C1MINPF = Character 1 Min Phoneme Frequency, C1MAXPF = Character 1 Max Phoneme Frequency, C1IPF = Character 1 Initial Phoneme Frequency, C1MINDF = Character 1 Min Diphone Frequency, C1IDF = Character 1 Initial Diphone Frequency.

predictor	C1MEPF	C1MINPF	C1MAXPF	C1IPF	C1MINDF	C1IDF
C1MEPF	-					
C1MINPF	0.560	-				
C1MAXPF	0.893	0.299	-			
C1IPF	0.489	0.870	0.293	-		
C1MINF	0.700	0.581	0.663	0.526	-	
C1IDF	0.699	0.573	0.676	0.539	0.975	

strong correlation between the initial and minimum phoneme frequency ($r = 0.870$) follows straightforwardly from the fact that for 24,836 words (81.04%) the initial phoneme of the first character is the least frequent phoneme in that character.

2.4.4.6 Cluster 14: character 2 phonological frequency

Cluster 14 is the last cluster of phonological measures and consists of 11 predictors. The first 6 measures are the character 2 counterparts of the numerical variables in Cluster 13: `Character 2 Mean Phoneme Frequency`, `Character 2 Min Phoneme Frequency`, `Character 2 Max Phoneme Frequency` and `Character 2 Initial Phoneme Frequency`, `Character 2 Min Diphone Frequency`, and `Character 2 Initial Diphone Frequency`. The next 4 measures describe the phonological frequency of the word as a whole: `Mean Phoneme Frequency`, `Min Phoneme Frequency`, `Max Phoneme Frequency`, and `Min Diphone Frequency`.

The last measure in Cluster 14 is `Transitional Diphone Frequency`. `Transitional Diphone Frequency` is the frequency of the diphone that connects the pronunciations of the first and the second character. The pronunciation of the word 玫瑰 ("rose"), for instance, is [meɪ2kui1]. The transitional diphone, therefore, is [ɪk]. This diphone has a frequency of 1058.12. `Transitional Diphone Frequency` for the word 玫瑰, thus, is 1058.12.

`Mean Diphone Frequency` and `Max Diphone Frequency` cluster with the corresponding measures for character 1. By contrast, `Min Diphone Frequency` clusters with character 2 diphone frequency measures. Table 2.19 shows the correlation of

Table 2.19: Correlations of word-level phonological frequency measures with the corresponding character 1 and character 2 phonological frequency measures. Maximum correlations for each word-level predictor are printed in bold font.

predictor	C1	C2
Mean Phoneme Frequency	**0.753**	0.694
Min Phoneme Frequency	**0.669**	0.604
Max Phoneme Frequency	**0.612**	0.611
Mean Diphone Frequency	**0.724**	0.664
Min Diphone Frequency	0.343	**0.434**
Max Diphone Frequency	**0.675**	0.609

each word-level phonological frequency measures with its character 1 and character 2 counterparts. All correlations are significant at the 0.001 α level. As before, the clustering algorithm follows the distributional properties of the data. While `Mean Diphone Frequency` and `Max Diphone Frequency` correlate more strongly with the character 1 measures, `Min Diphone Frequency` correlates more strongly with `Character 2 Min Diphone Frequency` than with `Character 1 Min Diphone Frequency`.

The clustering pattern for word-level phoneme frequency measures does not follow straightforwardly from the distributional structure of the data. For all three word-level phoneme frequency measures, the correlation with the corresponding character 1 measure is stronger than the correlation with the corresponding character 2 measure. Nonetheless, `Mean Phoneme Frequency`, `Min Phoneme Frequency`, and `Max Phoneme Frequency` all cluster with character 2 phoneme frequency measures, rather than with character 1 phoneme frequency measures. Alternative SOMs with the same structure but a different random initialization of weights typically showed the same clustering of word-level phoneme frequency measures with character 2 phoneme frequency measures.

A potential explanation for the fact that word-level phoneme frequency measures cluster with character 2 phoneme frequency measures, rather than with character 1 phoneme frequency measures comes from the spatial organization of the SOM (see Figure 2.3). Lexical properties of the second character are located at the top of the SOM, whereas lexical properties of the first character are located at the bottom of the SOM. Cluster 14 neighbours Cluster 8. Cluster 8 describes the visual complexity of the second character, as well as the visual complexity of the word as a whole. The presence of word-level phoneme frequency measures in Cluster 8 may thus be a consequence of local attraction between word-level predictors in the SOM.

Table 2.20: Pairwise Spearman correlations for the numerical variables in cluster 14. Abbreviations: C2MEPF = Character 2 Mean Phoneme Frequency, C2MAXPF = Character 2 Max Phoneme Frequency, C2MINPF = Character 2 Min Phoneme Frequency, C2IPF = Character 2 Initial Phoneme Frequency, C2MINDF = Character 2 Min Diphone Frequency, C2IDF = Character 2 Initial Diphone Frequency, MEPF = Mean Phoneme Frequency, MAXPF = Max Phoneme Frequency, MINPF = Min Phoneme Frequency, MINDF = Min Diphone Frequency, TDF = Transitional Diphone Frequency.

predictor	C2MEPF	C2MAXPF	C2MINPF	C2IPF	C2MINDF	C2IDF	MEPF	MAXPF	MINPF	MINDF	TDF
C2MEPF	-										
C2MAXPF	0.894	-									
C2MINPF	0.586	0.329	-								
C2IDF	0.509	0.308	0.863	-							
C1MINDF	0.701	0.679	0.583	0.524	-						
C1IDF	0.697	0.687	0.574	0.539	0.978	-					
MEPF	0.694	0.642	0.402	0.350	0.496	0.495	-				
MAXPF	0.344	0.218	0.604	0.506	0.383	0.378	0.445	-			
MINPF	0.553	0.611	0.198	0.189	0.424	0.428	0.759	0.167	-		
MINDF	0.363	0.299	0.441	0.450	0.434	0.440	0.485	0.491	0.233	-	
TDF	0.335	0.255	0.467	0.503	0.366	0.376	0.451	0.388	0.273	0.861	-

As can be seen in Table 2.20 all pairwise correlations in Cluster 14 are positive and significant at the 0.001 α level. The pairwise correlations for the character 2 phonological frequency measures in Cluster 14 and the corresponding measures in Cluster 13 are highly similar ($r = 0.999$). The distributional space for the phonological frequency measures for character 1 and character 2, thus, are nearly identical.

2.4.5 Group 4: homographs

Group 4 consists of 4 clusters that provide information about the orthography-to-phonology consistency of characters and phonetic radicals. These clusters are shown in purple in Figure 2.3. The first two clusters, Cluster 15 and Cluster 16, contain information about the type and token counts of character 1 and character 2 homographs, as well as about their frequency. The last two clusters, Cluster 17 and Cluster 18, consist of orthography-to-phonology consistency measures for the phonetic radicals of both characters.

2.4.5.1 Cluster 15: character 1 homographs

As can be seen in Table 2.21, Cluster 15 consists of 3 numerical variables related to the number and the frequency of homographs of the first character: `Character 1 Homographs (Types)`, `Character 1 Homographs (Tokens)`, and `Character 1 Homographs Frequency`. A homograph is a character that looks the same, but is pronounced differently. In other words: `Character 1 Homographs (Types)` is the number of different pronunciations for the first character. For example, the first character 差 of the word 差使 ("messenger"; "[tʰʂɑɪ1ʂi3]") is pronounced as "[tʰʂɑɪ1]". Apart from this pronunciation, the character 差 has 3 other pronunciations across the 30,645 words in the CLD: [tʰʂä1] (e.g., in 差数, "favourable balance", "[tʰʂä1ʂu4]"), [tʰʂä4] (e.g., in 利差, "interest margin", [li4tʰʂä4]), and [tʰsi1] (e.g., in 参差, "uneven", [tʰsən1tʰsi1]). `Character 1 Homographs Types` for the word 差使 therefore is 3.

Typically the number of homographs for a character is low. As can be seen in Table 2.21, the average number of character 1 homograph types for the words in the CLD is no greater than 0.23. Although homography is more common in Chinese than in English, it is therefore certainly not the case that each character in Chinese can be pronounced in many different ways.

Table 2.21: Overview of numerical predictors: homographs (Group 4)

	mean	median	sd	min	max	NA
Cluster 15: character 1 homographs						
C1 Homographs (Types)	0.23	0	0.50	0	3	0
C1 Homographs (Tokens)	1.79	0	10.64	0	366	0
C1 Homographs Freq.	45.04	0.00	477.33	0.00	45890.53	0
Cluster 16: character 2 homographs						
C2 Homographs (Types)	0.29	0	0.55	0	3	4710
C2 Homographs (Tokens)	4.20	0	21.93	0	366	4710
C2 Homographs Freq.	76.74	0.00	694.78	0.00	45938.18	4710
Cluster 17: character 1 phonetic radical enemies						
C1 PR Enemies (Types)	3.43	3	2.91	0	14	14478
C1 PR Enemies (Tokens)	47.12	26	55.56	0	302	14478
C1 PR Enemies Freq	767.93	150.12	2143.75	0.00	46504.14	14478
C1 PR Family Size	6.70	6	4.57	1	19	14478
C1 PR Frequency	976.60	449.70	1612.30	0.00	28622.72	14478
Cluster 18: character 2 phonetic radical enemies						
C2 PR Enemies (Types)	3.44	3	2.88	0	14	17944
C2 PR Enemies (Tokens)	46.60	25	54.78	0	302	17944
C2 PR Enemies Freq	886.95	153.82	2951.46	0.00	46501.34	17944
C2 PR Family Size	6.61	6	4.55	1	19	17944

`Character 1 Homographs Tokens` is the number of words in which the first character is pronounced differently. For the first character 差 in the word 差使, the alternative pronunciations [tʰʂä1], [tʰʂä4], and [tʰsi1] occur in 17, 4 and 1 words, respectively. Therefore, `Character 1 Homographs (Tokens)` for the word 差使 is $17 + 4 + 1 = 22$.

Finally, `Character 1 Homographs Frequency` is the summed frequency of all character 1 homograph tokens. The summed frequency of the 22 homograph tokens for the character 差 in the word 差使 is 247.56. `Character 1 Homograph Frequency` for the word 差使, therefore, is 247.56.

Table 2.22 presents the pairwise correlations between the measures in Cluster 15. All correlations are near-perfect and highly significant at the 0.001 α level. The 3 measures `Character 1 Homographs (Types)`, `Character 1 Homographs (Tokens)`, `Character 1 Homographs Frequency` thus encode very similar information.

Table 2.22: Pairwise Spearman correlations for the numerical variables in Cluster 15. Abbreviations: C1HTY = Character 1 Homographs (Types), C1HTO = Character 1 Homographs (Tokens), C1HF = Character 1 Homograph Frequency.

predictor	C1HTY	C1HTO	C1HF
C1HTY	-		
C1HTO	0.993	-	
C1HF	0.991	0.995	-

2.4.5.2 Cluster 16: character 2 homographs

Cluster 16 is the character 2 counterpart of Cluster 15 and contains 3 measures related to the number and frequency of homographs for character 2 (`Character 2 Homographs (Types)`, `Character 2 Homographs Tokens` and `Character 2 Homographs Frequency`). Table 2.23 presents the pairwise Spearman correlations for the measures in Cluster 16. Like the measures in Cluster 15, the numerical variables in Cluster 16 show near-perfect correlations that are highly significant at an α level of 0.001. Similar to the measures in Cluster 15, therefore, the variables in Cluster 16 encode highly similar information. The pairwise correlations for Cluster 15 and Cluster 16 correlate strongly ($r = 0.985$). The distributional spaces for the homography measures, thus, are highly similar for both characters.

Yet, there is a subtle but important difference between the homography measures for the first and the second character. The distribution of homography across characters seems fine-tuned to the information-theoretic properties of the immediate linguistic context in which a character appears. The means for all three character 2 measures (`Character 2 Homographs (Types)`: 0.29, `Character 2 Homographs (Tokens)`: 4.20, `Character 2 Homographs Frequency`: 76.74) are higher than the corresponding means for the character 1 measures for two-character words in Cluster

Table 2.23: Pairwise Spearman correlations for the numerical variables in Cluster 16. Abbreviations: C2HTY = Character 2 Homographs (Types), C2HTO = Character 2 Homographs (Tokens), C2HF = Character 2 Homograph Frequency.

predictor	C2HTY	C2HTO	C2HF
C2HTY	-		
C2HTO	0.985	-	
C2HF	0.983	0.993	-

15 (`Character 1 Homographs (Types)`: 0.25, `Character 1 Homographs (Tokens)`: 1.99, `Character 1 Homographs Frequency`: 51.25). As indicated by paired t-tests for the first and second character homography measures for all two-character words, these differences are significant (types: t(25934) = -9.89, $p < 0.001$; tokens: t(25934) = -14.47, $p < 0.001$; frequency: t(25934) = -4.74, $p < 0.001$). This suggests that the information provided by the first character reduces the uncertainty about the identity of the second character to such an extent that more variation is possible for the pronunciation of the second character.

Furthermore, characters that form single-character words (`Character 1 Homographs (Types)`: 0.11, `Character 1 Homographs (Tokens)`: 0.69, `Character 1 Homographs Frequency`: 10.86) show less homography than first characters in two-character words (`Character 1 Homographs (Types)`: 0.25, `Character 1 Homographs (Tokens)`: 1.99, `Character 1 Homographs Frequency`: 51.25). Again, these differences are significant (types: t(8805.69) = -22.67, $p < 0.001$; tokens: t(13322.63) = -12.18, $p < 0.001$; frequency: t(28174.99) = -10.85, $p < 0.001$). In the context of the information provided by a second character, therefore, the first character is allowed to provide less conclusive information about its pronunciation than when it appears by itself.

These observations suggest that when the uncertainty is sufficiently reduced, the phonological form of a character is allowed to vary. When it is not, a character is preferred to map onto a single phonological form. This fits well with discrimination learning approaches, in which the distributional properties of the language processing system are shaped by the need to reduce uncertainty about the linguistic input (see, e.g., Ramscar et al., 2013).

2.4.5.3 Cluster 17: character 1 phonetic radical orthography-to-phonology consistency

Clusters 15 and 16 contain numerical variables regarding the orthography-to-phonology consistency at the character level. The measures in Clusters 17 (character 1) and Cluster 18 (character 2) encode information about the orthography-to-phonology consistency of the phonetic radical. Cluster 17 consists of 5 measures. The first 3 measures are the counterparts of the measures in Cluster 15 at the phonetic radical level: `Character 1 PR Enemies (Types)`, `Character 1 PR Enemies (Tokens)`, and `Character 1 PR Enemies Frequency`.

`Character 1 PR Enemies (Types)` is the number of different pronunciations of characters in which the phonetic radical of the first character appears. For example, the phonetic radical of the first character 端 in the word 端倪 ("clue", "[tuän1ni2]") is 耑. In addition to "[tuän1]", there are 4 other pronunciations of characters that contain this phonetic radical: "[t^hʂuaɪ4]" (e.g., in 踹, "to kick", "[t^hʂuaɪ4]"), "[t^hʂuɑn3]" (e.g., in the first character 喘 of the word 喘息, "to pant", "[t^hʂuɑn3t^hɕi4]"), "[t^huɑn1]" (e.g., in the first character 湍 of the word 湍流, "rushing water", "[t^huɑn1liu2]"), and "[ʐ̩ui4]" (e.g., in 瑞, "propitious", "[ʐ̩ui4]"). Therefore, `Character 1 PR Enemies (Types)` for the word 端倪 is 4.

`Character 1 PR Enemies (Tokens)` refers to the number of words in which the character that has the same phonetic radical as the first character of the current word is pronounced differently than the first character in the current word. The phonetic radical 耑 is pronounced as "[t^hʂuaɪ4]" in 1 word, as "[t^hʂuɑn3]" in 5 words, as "[t^huɑn1]" in 3 words and as "[ʐ̩ui4]" in 2 words. `Character 1 PR Enemies (Tokens)` for the word 端倪, therefore, is $1 + 5 + 3 + 2 = 11$.

`Character 1 PR Enemies Frequency` is the summed frequency of the enemy tokens. The summed frequency of the 11 words that are phonetic radical enemies of the first character in the word 端倪 is 20.37. `Character 1 PR Enemies Frequency`, therefore, is 20.37.

Compared to the character-level orthography-to-consistency measures (`Character 1 Homographs (Types)`: 0.23, `Character 1 Homographs (Tokens)`: 1.79, `Character 1 Homographs Frequency`: 45.04), the corresponding phonetic radical measures have higher means (`Character 1 PR Enemies (Types)`: 3.43, `Character 1 PR Enemies (Tokens)`: 47.12, `Character 1 PR Enemies Frequency`: 767.93). Homography at the phonetic radical level, therefore, is much more common than at the character level.

The fourth measure in Cluster 17 is `Character 1 PR Family Size`, which is defined as the number of characters the phonetic radical of the first character occurs in. For the word 端倪, for instance, the phonetic radical 耑 of the first character 端, appears in 5 characters (端, 踹, 喘, 湍, 瑞). `Character 1 PR Family Size` for the word 端倪 thus is 5. The final numerical variable in Cluster 17 is `Character 1 PR Frequency`, which is the first character equivalent of `Character 2 PR Frequency` (see the discussion of Cluster 6).

Table 2.24: Pairwise Spearman correlations for the numerical variables in Cluster 17. Abbreviations: C1PRENTY = Character 1 PR Enemies (Types), C1PRENTO = Character 1 PR Enemies (Tokens), C1PRENFR = Character 1 PR Enemies Frequency, C1PRFS = Character 1 PR Family Size, C1PRF = Character 1 PR Frequency.

predictor	C1PRENTY	C1PRENTO	C1PRENFR	C1PRFS	C1PRF
C1PRENTY	-				
C1PRENTO	0.844	-			
C1PRENFR	0.752	0.924	-		
C1PRFS	0.847	0.795	0.684	-	
C1PRF	0.588	0.726	0.783	0.543	-

Table 2.24 presents the pairwise correlations for the variables in Cluster 17. All pairwise correlations are positive and significant at an α level of 0.001. The lowest pairwise correlation is the correlation between `Character 1 PR Frequency` and `Character 1 PR Family Size`. This correlation is no less than 0.543. Cluster 17, therefore, is a highly homogeneous cluster.

2.4.5.4 Cluster 18: character 2 phonetic radical orthography-to-phonology consistency

Cluster 18 is the character 2 counterpart of Cluster 17 and contains the numerical variables `Character 2 PR Enemies (Types)`, `Character 2 PR Enemies (Tokens)`, `Character 2 PR Enemies Frequency`, and `Character 2 PR Family Size`. The pairwise correlations for the numerical variables in Cluster 18 are presented in Table 2.25. All correlations are positive and significant at an α level of 0.001. Furthermore, the

Table 2.25: Pairwise Spearman correlations for the numerical variables in Cluster 18. Abbreviations: C2PRENTY = Character 2 PR Enemies (Types), C2PRENTO = Character 2 PR Enemies (Tokens), C2PRENFR = Character 2 PR Enemies Frequency, C2PRFS = Character 2 PR Family Size.

predictor	C2PRENTY	C2PRENTO	C2PRENFR	C2PRFS
C2PRENTY	-			
C2PRENTO	0.842	-		
C2PRENFR	0.749	0.913	-	
C2PRFS	0.837	0.793	0.671	-

pairwise correlations for Cluster 18 are highly similar to the pairwise correlations for the corresponding measures in Cluster 17 ($r = 0.998$). This indicates that the distributional structure of the phonetic radical orthography-to-consistency measures is similar for character 1 and character 2.

2.4.6 Group 5: homophones

The clusters in Group 4 contained measures describing the orthography-to-phonology consistency at the level of the character and the phonetic radical. The clusters in Group 5 describe consistency in the other direction: from phonology to orthography. As can be seen in Table 2.26, Cluster 19 describes the phonology-to-orthography consistency for character 1, both at the character-level and at the level of the phonetic radicals. Cluster 20 contains phonology-to-orthography measures for character 2.

2.4.6.1 Cluster 19: character 1 homophones

Cluster 19 consists of 6 phonology-to-orthography consistency measures for character 1. The first 3 measures encode information about the number of homophones

Table 2.26: Overview of numerical predictors: homophones (Group 5)

	mean	median	sd	min	max	NA
Cluster 19: character 1 homophones						
C1 Homophones (Types)	7.10	5	6.74	0	40	0
C1 Homophones (Tokens)	67.29	42	79.16	0	604	0
C1 Homophones Freq.	1378.86	358.68	2739.90	0.00	46911.48	0
C1 PR BE (Types)	2.91	2	3.10	0	16	14478
C1 PR BE (Tokens)	30.51	14	40.92	0	270	14478
C1 PR BE Freq.	477.66	43.31	1306.58	0.00	45723.29	14478
Cluster 20: character 2 homophones						
C2 Homophones (Types)	7.82	6	7.72	0	40	4710
C2 Homophones (Tokens)	77.40	46	95.24	0	604	4710
C2 Homophones Freq.	1771.20	426.66	3620.35	0.00	46911.48	4710
C2 PR BE (Types)	3.23	2	3.37	0	16	17944
C2 PR BE (Tokens)	34.66	19	43.87	0	269	17944
C2 PR BE Freq.	555.94	70.31	1149.36	0.00	45723.29	17944

of a character. Homophones are characters that are pronounced the same, but have a different orthographic form.

`Character 1 Homophones (Types)` is the number of orthographically distinct characters that have the same pronunciation as the current word-initial character. For example, the first character of the word 阻挡 ("to obstruct", "[tsu3tɑŋ3]"), 阻, is pronounced as "[tsu3]". There are 4 other characters that share this pronunciation: 组 (e.g., in 组成, "component"), 祖 (e.g., in 祖上, "ancestor"), 诅 (e.g. in 诅咒, "curse"), and 俎 (in 俎, "chopping board"). `Character 1 Homophones (Types)` for the word "阻挡", therefore, is 4.

`Character 1 Homophones (Tokens)` is the token count of characters that have the same pronunciation as the current word-initial character, but have a different orthographic form. The 4 characters 组, 祖, 诅 and 俎 are pronounced as "[tsu3]" in 22, 21, 2 and 1 words, respectively. `Character 1 Homophones (Tokens)` for the word 阻挡 therefore is $22 + 21 + 2 + 1 = 46$.

The third homophony measure for the first character is `Character 1 Homophones Frequency`. `Character 1 Homophones Frequency` is defined as the summed frequency of all character 1 homophone tokens. The summed frequency of the 22 homophones for the character 阻 is 1503.74. `Character 1 Homophones Frequency` for the word 阻挡 therefore is 1503.74.

As can be seen in Table 2.26, the average number of character 1 homophone types is 7.10 and the average number of homophone tokens is 67.29. Homophone counts for character 1 are thus much higher than homograph counts for character 1 (mean number of types: 0.23, mean number of tokens: 1.79).

The abundance of homophones is a result of the rich orthographic system in Chinese. In total, the $30,645$ words in the CLD contain $5,242$ unique characters. The number of phonological forms in Chinese is much more limited. The number of unique character pronunciations in the CLD, including tones, is 1,307. This implies that no more than 0.25 unique pronunciations are available per unique character, while 4.01 unique characters are available per unique pronunciation. Unlike homography, homophony is therefore ubiquitous in Mandarin Chinese, with a mere 7.83% of all words in the CLD having no character 1 homophones.

The other 3 variables in Cluster 19 are phonology-to-orthography consistency measures at the level of the phonetic radical. `Character 1 PR Backward Enemies (Types)` is the number of distinct phonetic radicals for which the character that contains it is pronounced in the same way as the first character of the current word.

The first character 福 ("[fu2]") in the word 福祉 ("welfare","[fu2ʈʂi3]"), for instance, contains the phonetic radical 畐. Characters that are also pronounced as "[fu2]", contain 4 other phonetic radicals: 孚 (e.g., in the first character of 浮筒, "buoy", "[fu2tʰɔŋ3]"), 付 (e.g., in the second character of 音符, "music note", "[in1fu2]"), 夫 (e.g., in the first character of 扶养, "to raise a child", "[fu2iɑŋ3]"), 弗 (e.g., in the second character of 仿佛, "as if", "[fɑŋ3fu2]"). `Character 1 PR Backward Enemies (Types)` for the word 福祉, therefore, is 4.

As before, `Character 1 PR Backward Enemies (Tokens)` is the token count for the phonetic radical backward enemies of the first character. The 4 phonetic radical backward enemies for the first character of the word 福祉 appear in 36 (孚), 9 (付), 12 (夫), and 7 (弗) words. `Character 1 PR Backward Enemies (Tokens)` for the word 福祉, hence, is $36 + 9 + 12 + 7 = 64$.

The final measure of phonology-to-orthography consistency at the phonetic radical level is `Character 1 PR Backward Enemies Frequency`, which is the summed frequency of the phonetic radical backward enemies tokens for the first character. For the word 福祉, `Character 1 PR Backward Enemies Frequency` is the summed frequency of the 64 backward enemies of the phonetic radical in the first character, which is 437.95.

The pairwise correlations for the variables in Cluster 19 are shown in Table 2.27. All correlations are positive and significant at an α level of 0.001. Cluster 19, therefore, is a homogeneous cluster of the phonology-to-orthography consistency of the first character.

Table 2.27: Pairwise Spearman correlations for the numerical variables in Cluster 19. Abbreviations: C1HPTY = Character 1 Homophones (Types), C1HPTO = Character 1 Homophones (Tokens), C1HPF = Character 1 Homophones (Frequency), C1PRBETY = Character 1 PR Backward Enemies (Types), C1PRBETO = Character 1 PR Backward Enemies (Tokens), C1PRBEF = Character 1 PR Backward Enemies (Frequency).

predictor	C1HPTY	C1HPTO	C1HPF	C1PRBETY	C1PRBETO	C1PRBEF
C1HPTY	-					
C1HPTO	0.840	-				
C1HPF	0.743	0.914	-			
C1PRBETY	0.850	0.714	0.628	-		
C1PRBETO	0.793	0.778	0.677	0.851	-	
C1PRBEF	0.719	0.727	0.707	0.791	0.939	-

2.4.6.2 Cluster 20: character 2 homophones

Cluster 20 is the character 2 counterpart of Cluster 19 and likewise consists of 6 numerical variables. As was the case for Cluster 19, the first 3 predictors are measures of the phonology-to-orthography consistency at the character level: `Character 2 Homophones (Types)`, `Character 2 Homophones (Tokens)`, and `Character 2 Homophones Frequency`. For the homography measures in Cluster 15 and Cluster 16, the means of the character 2 measures were higher than the means of the character 1 measures. For the homophony measures in Clusters 19 and Cluster 20 the same holds true. The differences between the means of the character-level homophony measures for the first character (`Character 1 Homophones (Types)`: 7.10, `Character 1 Homophones (Tokens)`: 67.29, `Character 1 Homophones Frequency`: 1378.86) and the means of the equivalent measures for the second character (`Character 2 Homophones (Types)`: 7.82, `Character 2 Homophones (Tokens)`: 77.40, `Character 2 Homophones Frequency`: 1771.20), however, are smaller than the differences for the homography measures.

Analogous to Cluster 19, the second set of 3 measures in Cluster 20 contains information about the phonology-to-orthography consistency at the level of the phonetic radical. The numerical variables in this set are `Character 2 PR Backward Enemies (Types)`, `Character 2 PR Backward Enemies (Tokens)`, and `Character 2 PR Backward Enemies Frequency`.

Table 2.28 provides the pairwise correlations between the predictors in Cluster 20. All pairwise correlations for Cluster 19 were highly significant. Likewise, all pairwise correlations for Cluster 20 are significant at the 0.001 α level. The pairwise correlations for Cluster 19 and 20 are nearly identical ($r = 0.993$). As was the case for the frequency, entropy, visual complexity, phonological frequency and homography measures, therefore, the distributional space for the character 1 and character 2 measures of homophony is highly similar. The similarity of the pairwise correlations for character 1 and character 2 measures in the CLD suggests that lexical properties are encoded in much the same way for the first and second character.

2.4.7 Group 6: other predictors

The final group of clusters, Group 6, consists of a single cluster (Cluster 21). This cluster contains a variety of numerical predictors that – according to the clustering algorithm we applied to the SOM – did not fit into one of the 20 clusters discussed above. The numerical variables in Group 6 are presented in Table 2.29.

Table 2.28: Pairwise Spearman correlations for of the numerical variables in Cluster 20. Abbreviations: C2HPTY = Character 2 Homophones (Types), C2HPTO = Character 2 Homophones (Tokens), C2HPF = Character 2 Homophones (Frequency), C2PRBETY = Character 2 PR Backward Enemies (Types), C2PRBETO = Character 2 PR Backward Enemies (Tokens), C2PRBEF = Character 2 PR Backward Enemies (Frequency).

predictor	C2HPTY	C2HPTO	C2HPF	C2PRBETY	C2PRBETO	C2PRBEF
C2HPTY	-					
C2HPTO	0.849	-				
C2HPF	0.745	0.917	-			
C2PRBETY	0.841	0.698	0.613	-		
C2PRBETO	0.775	0.774	0.674	0.835	-	
C2PRBEF	0.704	0.735	0.720	0.769	0.936	-

2.4.7.1 Cluster 21: other measures

As can be seen in Figure 2.3, Cluster 21 spans a large part of the SOM (19 of the 100 output units). Correspondingly, it is the cluster that contains the largest number of numerical variables. In total, Cluster 21 contains no less than 24 predictors. An inspection of the measures in Cluster 21 revealed that 5 different types of information are encoded by the numerical variables in this cluster. We added dashed lines to Table 2.29 to separate the different types of predictors.

The first type of predictors encodes the frequency of a word and its characters in traditional Chinese: `Traditional Frequency`, `Character 1 Traditional Frequency`, and `Character 2 Traditional Frequency`. The frequency measures (per million) were obtained from the Academia Sinica Corpus (henceforth AS corpus; Academia Sinica, 1998), which is a balanced corpus of traditional Chinese that consists of 5.4 million words and 8.4 million characters. The frequency counts per million for the characters and words in the CLD are lower in the AS corpus than in the three corpora of simplified Chinese. This is the case at the word level (mean frequency AS: 15.70; mean frequencies SCCoW, Gigaword and SUBTLEX-CH: 22.68, 22.20, and 30.28)[5], as well as for character 1 (mean frequency AS: 524.07; mean frequencies SCCoW, Gigaword and SUBTLEX-CH: 688.16, 694.38, and 676.59) and character 2 (mean frequency AS: 558.28; mean frequencies SCCoW, Gigaword and SUBTLEX-CH: 783.67, 760.61, and 762.00).

[5]The mean word frequency is higher for SUBTLEX-CH than for the SCCoW and Gigaword corpora. This may be a consequence of using the presence of a word in the SUBTLEX-CH word frequency list as one of the criteria for including a word in the CLD.

Table 2.29: Overview of numerical predictors: other predictors (Group 6). Phonological frequencies were rounded to 1 decimal place to prevent the table from exceeding the page width. Dashed lines are added for ease of interpretation only and do not correspond to clusters in the hierarchical clustering analysis carried out on the SOM.

	mean	median	sd	min	max	NA
Cluster 21: other measures						
Traditional Chinese Freq.	15.70	0.00	356.47	0.00	52041.78	0
Trad. Chin. C1 Freq.	524.07	43.50	1349.82	0.00	34628.50	0
Trad. Chin. C2 Freq.	558.28	71.10	1288.99	0.00	34628.50	4710
C1 Mean HLC Freq.	22487.9	19325.9	15504.3	0.0	101583.1	3
C2 Mean HLC Freq.	22842.2	18980.1	16344.4	0.3	101583.1	4711
C1 Min HLC Freq.	8264.9	5234.6	10040.8	0.0	101583.1	3
C2 Min HLC Freq.	8965.5	5973.1	11016.9	0.3	101583.1	4711
C1 Max HLC Freq.	40311.4	30409.3	30277.7	0.0	101583.1	3
C2 Max HLC Freq.	39963.4	30356.7	30674.1	0.3	101583.1	4711
C1 Mean LLC Freq.	236140.6	238706.4	80796.1	1.8	589864.0	3
C2 Mean LLC Freq.	235178.7	237335.2	78512.3	387.3	589864.0	4711
C1 Min LLC Freq.	56305.1	29018.7	61593.3	0.1	589864.0	3
C2 Min LLC Freq.	57647.5	31523.3	59017.3	30.7	589864.0	4711
C1 Max LLC Freq.	475638.7	589864.0	151246.0	1.8	589864.0	3
C2 Max LLC Freq.	470827.9	589864.0	151600.1	387.3	589864.0	4711
C1 SR Frequency	11882.48	7579.14	10666.60	0.00	36508.12	0
C2 SR Frequency	11398.76	7579.14	10198.72	1.91	36508.12	4710
C1 SR Family Size	85.59	49	89.26	1	286	0
C2 SR Family Size	79.21	40	84.24	1	286	4710
C1 SR Strokes	3.66	3	1.70	1	16	0
C2 SR Strokes	3.67	3	1.63	1	14	4710
C1 Relative Entropy	6.32	6.45	1.59	2.08	11.02	4713
C2 Relative Entropy	6.30	6.39	1.52	2.09	10.96	4711
Entropy Char. Freq.	0.62	0.68	0.32	0.00	1.00	4710

Furthermore, despite the medium-strength correlations between the traditional frequency measures from the AS corpus and the simplified frequency measures from the SCCOW (`Character 1 Frequency` and `Character 1 Traditional Frequency`: $r = 0.459$; `Character 2 Frequency` and `Character 2 Traditional Frequency`:

$r = 0.443$; `Frequency` and `Traditional Frequency`: $r = 0.226$), information about the traditional frequency measures is encoded by entirely different output neurons in the SOM than information about simplified frequency measures. As a result, simplified and traditional frequency measures are placed in different clusters in the agglomerative clustering technique applied to this SOM. These observations highlight the importance of using texts written in simplified Chinese to obtain accurate frequency counts for simplified Chinese.

The pairwise correlations for the traditional frequency measures in Cluster 21 are shown in Table 2.30. Both character frequency measures show positive correlations with word frequency that are significant at the 0.001 α level. The correlation between character 1 frequency and character 2 frequency is significant at an α level of 0.001 as well. The strength of this correlation, however, is weak ($r = 0.027$).

Table 2.30: Pairwise Spearman correlations for the traditional frequency measures in Cluster 21. Abbreviations: TF = Traditional Frequency, C1TF = Character 1 Traditional Frequency, C2TF = Character 2 Traditional Frequency.

predictor	TF	C1TF	C2TF
TF	-		
C1TF	0.479	-	
C2TF	0.490	0.027	-

The second type of numerical variables in Cluster 21 describes the frequency of the high-level and low-level components described in the discussion of Cluster 7 and Cluster 8 above. In total there are 12 numerical variables of this type. The first 6 measures describe the mean, minimum and maximum frequency of the high-level components in both characters (`Character 1 Mean High-Level Component Frequency`, `Character 1 Min High-Level Component Frequency`, `Character 1 Max High-Level Component Frequency`, `Character 2 Mean High-Level Component Frequency`, `Character 2 Min High-Level Component Frequency`, `Character 2 Max High-Level Component Frequency`). As mentioned in our discussion of Cluster 7, the first character 欣 of the word 欣喜 ("happy") consists of two high-level components: 斤 and 欠. The frequency of a component is defined as the summed frequency of all words that contain that component. The frequency of the high-level component 斤 is 8093.22, whereas the frequency of the high-level component 欠 is 6724.14. For the word 欣喜, therefore, `Character 1 Max High-Level Component` is 8092.96, `Character 1 Min High-Level Component Frequency` is 6567.51 and

`Character 1 Mean High-Level Component Frequency` is $\frac{8092.96+6567.51}{2} = 7330.24$. The character 2 counterparts of these measures are calculated analogously.

The other 6 measures are the low-level component counterparts of the high-level component frequency measures: `Character 1 Mean Low-Level Component Frequency`, `Character 1 Min Low-Level Component Frequency`, `Character 1 Max Low-Level Component Frequency`, `Character 2 Mean Low-Level Component Frequency`, `Character 2 Min Low-Level Component Frequency`, and `Character 2 Max Low-Level Component Frequency`. The first character 欣 of the word 欣喜 consists of 4 low-level components: "厂" (frequency: 15820.74), "𠂉" (frequency: 18,725.72), "𠂇" (frequency: 19038.38), and "人" (frequency: 132,942.50). For the word 欣喜, therefore, `Character 1 Max Low-Level Component Frequency` is 132038.14, `Character 1 Min Low-Level Component Frequency` is 15818.99, and `Character 1 Mean Low-Level Component Frequency` is $\frac{15820.74+18,725.72+19038.38+132,942.50}{4} = 46355.56$. The character 2 low-level component frequency measures are calculated in the same manner.

The pairwise correlations for the component frequency measures are shown in Table 2.31. Correlations that are not significant at the 0.001 α level are indicated with exclamation marks. All pairwise correlations between character 1 high-level component frequency measures are positive and significant at an α level of 0.001. The same holds for all pairwise correlations between character 2 high-level component frequency measures.

For the low-level components, mean frequency shows solid positive correlations with minimum and maximum frequency for both characters. The correlations between minimum and maximum low-level component frequency are also significant at the 0.001 α level, but are relatively weak (`Character 1 Min Low-Level Component Frequency` and `Character 1 Min Low-Level Component Frequency`: $r = 0.058$; `Character 2 Min Low-Level Component Frequency` and `Character 2 Min Low-Level Component Frequency`: $r = 0.034$).

Furthermore, most character 1 measures of component frequency do not correlate significantly (at the 0.001 α level) with character 2 measures of component frequency, and vice versa. The correlations between character 1 and character 2 measures that do reach significance are weak, with a maximum correlation between character 1 and character 2 component frequency measures of 0.036 for the correlation between `Character 1 Min Low-Level Component Freq` and `Character 2 Min Low-Level Component Freq`.

Table 2.31: Pairwise Spearman correlations for the component frequency measures in cluster 21. Exclamation marks indicate that a correlation did not reach significance at the 0.001 α level. Abbreviations: C1MEH = Character 1 Mean High-Level Component Frequency, C2MEH = Character 2 Mean High-Level Component Frequency, C1MIH = Character 1 Min High-Level Component Frequency, C2MIH = Character 2 Min High-Level Component Frequency, C1MAH = Character 1 Max High-Level Component Frequency, C2MAH = Character 2 Max High-Level Component Frequency, C1MEL = Character 1 Mean Low-Level Component Frequency, C2MEL = Character 2 Mean Low-Level Component Frequency, C1MIL = Character 1 Min Low-Level Component Frequency, C2MIL = Character 2 Min Low-Level Component Frequency, C1MAL = Character 1 Max Low-Level Component Frequency, C2MAL = Character 2 Max Low-Level Component Frequency.

predictor	C1MEH	C2MEH	C1MIH	C2MIH	C1MAH	C2MAH	C1MEL	C2MEL	C1MIL	C2MIL	C1MAL	C2MAL
C1MEH	-											
C2MEH	0.023	-										
C1MIH	0.532	$0.014^{!}$	-									
C2MIH	$0.011^{!}$	0.545	0.036	-								
C1MAH	0.948	0.021	0.327	$-0.000^{!}$	-							
C2MAH	0.021	0.953	$0.002^{!}$	0.344	0.023	-						
C1MEL	0.242	$0.012^{!}$	0.165	$0.018^{!}$	0.218	$0.005^{!}$	-					
C2MEL	$0.009^{!}$	0.248	$0.019^{!}$	0.189	$0.004^{!}$	0.225	0.024	-				
C1MIL	0.256	$-0.002^{!}$	0.377	$0.019^{!}$	0.144	$-0.011^{!}$	0.400	$0.020^{!}$	-			
C2MIL	$0.008^{!}$	0.241	0.027	0.421	$0.003^{!}$	0.124	$0.015^{!}$	0.391	0.026	-		
C1MAL	0.133	$-0.000^{!}$	−0.028	$-0.003^{!}$	0.171	$-0.000^{!}$	0.704	$0.005^{!}$	0.058	$0.002^{!}$	-	
C2MAL	$-0.002^{!}$	0.109	$0.000^{!}$	−0.030	$-0.002^{!}$	0.149	$0.005^{!}$	0.701	$0.005^{!}$	0.034	$0.003^{!}$	-

Despite being conceptually similar, the set of component frequency measures, therefore, is characterized by much more heterogeneity than the groups and clusters of conceptually similar numerical variables discussed above. This may explain why these measures are part of Cluster 21, rather than forming a separate component frequency cluster.

The third type of numerical variables in Cluster 21 provides lexical characteristics of the semantic radical. As demonstrated above, measures related to the phonetic radical either cluster with corresponding character-level measures (strokes, homophones) or form separate clusters in a group of clusters that contains corresponding character-level measures (homographs). Semantic radical measures, by contrast, cluster with conceptually unrelated measures in Cluster 21. The fact that phonetic radical measures, but not semantic radical measures cluster with character-level measures indicates that lexical properties of the phonetic radical are much more similar to lexical properties at the character-level than are lexical properties of the semantic radical.

In total, the CLD contains 6 numerical variables related to semantic radicals. For both characters, a measure of the frequency (`Character 1 SR Frequency`, `Character 2 SR Frequency`), the family size (`Character 1 SR Family Size`, `Character 2 SR Family Size`), and the number of strokes (`Character 1 SR Strokes`, `Character 2 SR Strokes`) of the semantic radical is provided. The frequency of a semantic radical is calculated in the same manner as the frequency of the phonetic radical. The semantic radical for the word 版权 ("copyright") is 片 ("piece"). This semantic radical occurs in 5 different characters: 版 (frequency: 252.14), 牌 (frequency: 354.62), 牍 (frequency: 0.68), 牒 (frequency: 0.61), and 片 (frequency: 303.65).[6] `Character 1 SR Frequency` for the word 版权 ("copyright") is the summed frequency of these 5 characters, which is 911.69.

The family size of the semantic radical, too, is defined analogously to the family size of the phonetic radical. As noted above, the semantic radical 片 of the first character in the word 版权 occurs in 5 characters. `Character 1 SR Family Size` for the word 版权, therefore is 5. `Character 1 SR Strokes` refers to the number of strokes in the semantic radical of the first character. The semantic radical 片 consists of 4 strokes. `Character 1 SR Strokes` for the word 版权 is 片, therefore, is 4. The semantic radical measures for the second character are calculated analogously to the semantic radical measures for the first character.

[6]Note that the character 片 is identical to its semantic radical 片. The ability of semantic radicals to function as independent characters is common, with a majority of semantic radicals appearing as independent characters as well.

Table 2.32: Pairwise Spearman correlations for the semantic radical measures in Cluster 21. Abbreviations: C1SRF = Character 1 SR Frequency, C2SRF = Character 2 SR Frequency, C1SRFS = Character 1 SR Family Size, C2SRFS = Character 2 SR Family Size, C1SRS = Character 1 SR Strokes, C2SRS = Character 2 SR Strokes.

predictor	C1SRF	C2SRF	C1SRFS	C2SRFS	C1SRS	C2SRS
C1SRF	-					
C2SRF	0.061	-				
C1SRFS	0.740	0.022	-			
C2SRFS	0.026	0.754	0.045	-		
C1SRS	−0.474	−0.049	−0.193	$-0.000^{!}$	-	
C2SRS	−0.035	−0.472	$0.001^{!}$	−0.201	0.056	-

Table 2.32 presents the pairwise correlations for the semantic radical measures. For both characters, the semantic radical frequency, family size and strokes measures are significantly correlated at the 0.001 α level, with the strongest correlations between the frequency and family size measures and the weakest correlations between the family size and strokes measures. As was the case for the traditional frequency and component frequency measures, correlations between the semantic radical measures of character 1 and character 2 are weak, with a maximum correlation of 0.061 between `Character 1 SR Frequency` and `Character 2 SR Frequency`.

The fourth type of lexical variable in Cluster 21 encodes the relative entropy of the first and second character (`Character 1 RE`, `Character 2 RE`). The relative entropy of two probability distributions, also known as the Kullback-Leibler divergence between the distributions, is defined as:

$$\sum_{i=1}^{n} p_i * \log_2(\frac{p_i}{q_i}) \tag{2.4}$$

For `Character 1 RE`, we defined the reference distribution q as the probability distribution of *second* characters across all two-character words in the CLD. For a given initial character, p was defined as the probability distribution of second characters for that character in word-initial position. For `Character 2 RE`, the reference distribution q is the probability distribution of *first* characters across all two-character words in the CLD. For a given second character, p is the probability distribution of first characters for the character in word-final position. `Character`

Table 2.33: Relative entropy: fictive example

word	freq.	prob. p	2nd character	freq.	prob. q
天气 (“weather”)	59	0.68	气	440	0.15
天使 (“angel”)	12	0.14	使	817	0.28
天际 (“skyline”)	7	0.08	际	868	0.29
天上 (“heaven”)	6	0.07	上	201	0.07
天职 (“duty”)	2	0.02	职	211	0.07
天才 (“genius”)	1	0.01	才	406	0.14

`1 RE` thus is identical for all two-character words with the same first character. Similarly, `Character 2 RE` is identical for all two-character words with the same second character. To avoid taking the logarithm of 0, we added the minimum observed frequency (0.0021) to all frequency counts prior to converting the frequency distributions to probability distributions.

Consider the fictive example for `Character 1 RE` in Table 2.33. The lexicon for this example contains 6 characters that occur as a second character. To calculate `Character 1 RE` for the character 天 (“sky”) we need two sets of frequencies. First, the frequencies of the 6 two-character words in which 天 is the first character are required. Second, we need the frequencies of the 6 second characters across all first characters. Converting both frequency distributions to probabilities yields the probability distributions p (the probability distribution of second characters for the first character 天) and q (the probability distribution of second characters for all first characters). To calculate `Character 1 RE` for the character 天, p and q are entered into Equation 2.4. For our example, this yields a relative entropy of 1.12.

The more similar the probability distributions p and q, the smaller the relative entropy. A small value for relative entropy therefore indicates that a first or second character combines with second or first characters in a typical way, whereas a large value for relative entropy indicates that a first or second character combines with second or first characters in an atypical way. At $r = 0.111$, the correlation between `Character 1 RE` and `Character 2 RE`, although significant at an α level of 0.001, is weak. The relative entropy for both characters, therefore, is relatively independent.

Finally, Cluster 21 contains a numerical variable that is defined as the entropy of the character frequencies in a two-character word: `Entropy Character Frequencies`. For the word 鲨鱼 (“shark”), the frequency of the first character 鲨 is 2.15, and the frequency of the second character is 鱼 104.08. Converting these frequencies

to probabilities gives a probability of 0.02 for the first character 鲨 and a probability of 0.98 for the second character 鱼. `Entropy Character Frequencies` for the word 鲨鱼, therefore, is $-\sum_{i=1}^{n} p_i * \log_2(p_i) = 0.14$. By definition, `Character 1 RE`, `Character 2 RE`, and `Entropy Character Frequencies` cannot be calculated for single character words. For single character words, therefore, these measures were set to "NA".

2.5 Online interface

The CLD is publicly available at `http://www.chineselexicaldatabase.com` and released under the GNU General Public License. As noted above, access to `http://www.chineselexicaldatabase.com` is password-protected until this dissertation is published. The password is `75090246`. The online interface to the CLD provides two options to access the data in the CLD. First, the database can be downloaded in `.txt` and `.csv` format (30.5 MB, zipped: 11.9 MB). In addition, the database is available as a data frame for the statistical software R (8.8 MB).

Second, the CLD can be accessed through a search interface. Users have the option to search the full database, or to submit lists of words, characters or radicals for which lexical information should be shown. Similarly, either the full set of variables can be shown or a user may select a subset of variables in which she is interested. For the categorical variables that describe the structure, type, and tone of a character, factor levels that should be included in the output can be manually selected (by default all factor levels are included). For numerical variables, minimum and maximum values can be set to limit the range of a variable in the output. The result of the search interface can be viewed in the browser or e-mailed to the user.

2.6 Conclusions

Lexical databases provide information about the distributional properties of a language and help carry out psycholinguistic studies in an efficient, yet thorough manner. Large-scale lexical databases have recently become available for a number of languages, including English (Coltheart, 1981), German (Heister et al., 2011), French (New et al., 2007). Here, we presented a lexical database for simplified Chinese that is substantially larger than existing resources (Y. Liu et al., 2007; Cai & Brysbaert, 2010; Sze et al., 2014). The database, the Chinese Lexical Database

(CLD), contains 141 numerical predictors and 23 categorical variables for 30,645 one-character words and two-character words. The CLD is publicly available and can be downloaded and searched at `http://www.chineselexicaldatabase.com`.

The lexical predictors in the CLD describe frequency and information-theoretic measures, as well as orthographic and phonological properties at the word level, the character level and the radical level. Frequency-related predictors include frequency and contextual diversity measures from three different corpora of simplified Chinese. One of these corpora is new large-scale corpus of web pages, the Simplified Chinese Corpus of Webpages (SCCoW). The frequency counts from the SCCoW are the basis for a number of information-theoretic measures in the CLD that are related to the combinatorial properties of characters, including association measures and entropy measures.

Orthographic measures include not only stroke counts, but also a wide variety of visual complexity and orthographic neighbourhood density measures at different grain sizes, ranging from pixels to visual components at or just below the radical level. Phonological measures include lexical predictors describing the frequency, the complexity and the neighbourhood density of character and word pronunciations. The CLD furthermore contains a wide range of predictors related to the orthography-to-phonology and phonology-to-orthography consistency, both for characters and for phonetic radicals.

We discussed the numerical predictors in the CLD on the basis of the results from a hierarchical clustering technique applied to the weights matrix of a self-organizing map (SOM; Kohonen, 1982) trained on the squared correlation matrix for these predictors. We highlighted key aspects of the correlational structure of the predictors in the CLD and observed a number of interesting patterns in the distributional lexical space for simplified Chinese. Frequency counts for traditional and simplified Chinese, for instance, were represented by entirely different neurons in the SOM and thus capture different information. Therefore, it is pivotal to use appropriate frequency counts when studying simplified or traditional Chinese.

Another interesting observation was that the consistency between orthography and phonology was greater for the first character than for the second character: second characters had more and more frequent homographs and homophones than first characters. For first characters, the uncertainty about the pronunciation of the character is relatively high. For second characters, by contrast, the amount of uncertainty about the pronunciation is substantially reduced through the informa-

tion provided by the first character. The extent to which the mapping between orthography and phonology varies thus is inversely proportional to the amount of uncertainty about the pronunciation of a character. Hence, the distributional properties of Chinese are shaped by the need to reduce uncertainty about the linguistic input (c.f., Ramscar et al., 2013).

In the rest of this dissertation we explore the explanatory power of the categorical and numerical variables in the CLD for experimental data. Chapter 3 presents an experiment in which a native reader of simplified Chinese was presented with all 30,645 words in the CLD in a word naming task. We discuss the effects of the lexical predictors in the CLD on naming latencies, pronunciation durations and eye fixation durations. The results for this experiment reveal interesting insights into lexical processing at and *below* the word level. Chapter 4 we gauge lexical processing at and `above` the word level through a phrase reading experiment. Based on the results of an analysis of the eye movement patterns we discuss how readers of simplified Chinese rapidly and efficiently integrate words to allow for an understanding of larger linguistic elements.

3
Word naming

3.1 Introduction

Chapter 2 introduced a new large-scale lexical database for simplified Chinese: the Chinese Lexical Database (CLD). This lexical database consists of a large number of categorical and numerical lexical variables. The question, then, arises to what extent and how these lexical variables can help predict dependent variables in psycholinguistic data sets. To be able to start answering this question, experimental data for the 30,645 words in the CLD are needed.

Different options are available with respect to the type of experimental data that would allow for an interesting first exploration of the predictive power of the lexical variables in the CLD. First, an experimental paradigm has to be chosen. The two most basic, well-established tasks in the psycholinguistic literature are lexical decision and word naming. The English Lexicon Project (henceforth ELP; Balota et al., 2007), for instance, provides both lexical decision latencies and word naming latencies for a large set of English words.

Lexical decision, however, is not as straightforward in Chinese as it is in English. One immediate question that comes to mind is how to construct items that are expected to yield a "no"-response, both at the word and at the character level. Chinese characters consists of multiple levels of information, ranging from the stroke to the radical level. Non-characters could be created by re-arranging or replacing elements at each of these levels. Sze et al. (2014), for instance, created non-characters for their single-character lexical decision experiment by replacing the semantic radical in a character with a semantic radical from a different character. Similarly,

non-characters could be created by re-arranging or replacing elements at the stroke level. At the word level, characters and non-characters could be combined in different ways to create non-words. One option would be to construct non-words that consist of non-characters only. Alternatively, either the left character, the right character, or both characters in two-character words could be non-characters. Another option would be to construct non-words by combining existing characters in an illegal manner. However, this would result in a different nature of "no"-stimuli for one-character words and for two-character words.

Decisions with respect to the construction of non-words and non-characters influence behavioural responses. For instance, the level at which non-character status is defined influences the amount of attention paid to the information at that level. Similarly, the nature of "no"-stimuli at the word level likely affects the nature of lexical processing.

To avoid these issues with the lexical decision task in Chinese, we collected experimental data for the words in the CLD using the word naming paradigm. The word naming paradigm allows for a stimulus list that consists entirely of valid words. Furthermore, the word naming paradigm provides the possibility to collect not only reaction times, but to also record pronunciations. In a first exploration of the data described below, we use these recordings for an analysis of the pronunciation durations. To obtain an even more thorough understanding of the processes that drive lexical processing in the word naming task we furthermore recorded the eye-movement patterns during the word naming task. Unlike reaction times or pronunciation durations, the eye-movement patterns provide temporal information that allows us to investigate the effects of different lexical variables over time. Taken together, naming latencies, pronunciation durations and eye fixation patterns provide a comprehensive overview of lexical processing during the word naming task.

A second issue related to the collection of experimental data concerns the selection of participants. Ideally, we would collect naming latencies for all $30,645$ words in the CLD for a large number of participants, representing both genders, different age groups and with different education levels. For the purpose of this dissertation, however, collecting thousands of hours of experimental data was infeasible. An alternative would be to collect data for a subset of the $30,645$ and a subset of the population of interest. Most typically, this would involve running a medium-scale experiment on several dozens of university students. Even among university students, however, there is considerable variation in language experience and so-

cial and geographical background between the native readers of simplified Chinese available in the Tübingen area.

Following Pham (2014), we therefore opted for a different approach here. Rather than presenting subsets of the CLD to different participants, we presented a single participant with all 30,645 words in the CLD. An obvious disadvantage of this approach is that generalizations to the general population of speakers of simplified Chinese or, more realistically, to a subset of this population are not possible. As noted by Pham (2014), however, a single-subject study has the advantage of reduced variance in the behavioural responses and therefore increased statistical power as compared to a study with multiple participants. After discussing the results of a large-scale single-participants study, Pham (2014, p. 114) concludes that "[...] the data obtained from a single, dedicated subject in a mega-study are of at least the same, if not higher quality compared to the data offered by a multi-participant study with informants with roughly the same general level of education".

Below, we first describe the single-participant word naming study for all words in the CLD in more detail. Next, we discuss the results of this study for three behavioural measures of language processing: naming latencies, pronunciation durations and eye fixation durations. For each behavioural measure we carry out an analysis using gradient boosting machines (GBMs) to investigate the quantitative properties of the effects of the lexical variables in the CLD and an analysis using generalized additive models (GAMs) to look at the qualitative nature of these effects.

3.2 Methods

3.2.1 Participants

One participant took part in the experiment. The participant is a 30-year-old male native reader of Mandarin Chinese, who was born in mainland China, lives in Tübingen and has corrected to normal vision.

3.2.2 Materials

The stimulus list for the experiment consisted of a list of 30,831 words that were present in the SUBTLEX-CH word frequency list (Cai & Brysbaert, 2010), as well as in the Contemporary Chinese Dictionary (Xiandai Hanyu Cidian, Chinese Academy

of Social Sciences, 2012). This list formed the basis for the 30,645 words in the CLD (see Section 2.2).

3.2.3 Design

The experiment consisted of 30,831 items that were randomly assigned to 31 experimental lists. The first 30 lists consisted of 1,000 items, whereas the last list consisted of 831 items. The order of items within a list was randomized.

Three behavioural measures were recorded: naming latencies, pronunciation durations and eye movement fixation durations. Naming latencies refer to the time in milliseconds from stimulus onset to pronunciation onset, whereas pronunciation durations denote the time in milliseconds from pronunciation onset to pronunciation offset. On the basis of Box-Cox tests, naming latencies were inverse transformed ($f(x) = \frac{-1000}{x}$) prior to analysis, whereas eye movement fixation durations were log-transformed. No transformation was necessary for pronunciation durations.

3.2.4 Procedure

The experiment was carried out in a soundproof booth. Naming latencies and acoustic durations were extracted from the recorded speech signal with the use of custom computer code using volume thresholds. The performance of this code was inspected on a trial-by-trial basis and corrected manually where necessary.

Eye movements were recorded with an EyeLink 1000 system, using a temporal resolution of 1,000 Hz. The experiment was run on a 17 inch LCD monitor using a 1,680 by 1,050 pixel resolution. The head of the participant was positioned on a chin rest that was located at a distance of 75 cm from the monitor. A 9-point grid calibration was carried out prior to the experiment, as well as after every 50 trials. The participant was instructed to limit eye blinking to a minimum during trials and to respond as fast as possible, while retaining accuracy.

Prior to each trial a fixation mark was shown in the center of the screen. When the participant fixated on this mark, a word was presented in the center of the screen in black SimHei 80 point font (i.e., the center of the word corresponded to the position of the fixation mark). The word remained on the screen for 2,000 milliseconds. After each stimulus, a blank screen appeared for 750 ms, followed by the fixation mark for the next trial. Each experimental session of 1,000 words had a duration of about 2 hours, including setup, calibrations and a 10 minute break halfway through the session.

3.3 Analysis

As mentioned above, three dependent variables were recorded during the experiment: naming latencies, pronunciation durations and fixation durations of the eye. For each of these dependent variables we carried out two types of analyses: a gradient boosting machine (GBM) and a generalized additive model (GAM). GBMs help provide insight into the relative influence of predictors on a dependent variable, whereas GAMs allow for an inspection of the qualitative nature of predictor effects.

3.3.1 Gradient boosting machines

Gradient boosting machines (GBMs; J. H. Friedman, 2001, 2002) combine the statistical concepts of gradient descent and boosting as a methodological advancement over standard tree-based ensemble methods, such as random forests. Tree-based ensemble methods are based on decision trees. Decision trees try to predict a dependent variable through a sequence of binary splits that lead to an optimal reduction in uncertainty about the value of the dependent variable.

In random forests (Strobl et al., 2009), a series of decision trees is combined to create an ensemble of trees: a forest. If each tree was fitted on the full data set, all trees in the forest would be identical and little would be gained by growing a forest. Therefore, the trees in random forests are grown on different subsets of the data. Subsetting is done both with respect to the data points, as well as with respect to the predictors.

For data points, subsets are created through a technique called bagging (bootstrap aggregating), with each tree being grown on a sample with replacement from the observations in the full data set. This sample has the same size as the full data set. Typically, around $\frac{2}{3}$ of the observations are in a bootstrap sample, while $\frac{1}{3}$ of the observations are not. The predictions of a tree for the observations that were not in the bootstrap sample can be used to obtain estimates of the prediction error for unseen data. The variance of the dependent variable in a collection of trees fitted to bootstrap samples is equal to the variance of individual trees divided by the number of trees. In other words: bagging leads to a reduction of the variance in the estimate of the dependent variable as compared to fitting a single decision tree.

However, the reduction of the variance in the estimate of the dependent variable is greatest when trees are less correlated. Bagging leads to trees with relatively high correlations. Thus, random forests use a subset not only the data points that serve

as input to a decision tree, but also of the predictors that are considered for each split (typically set to the square root of the total number of predictors in the data set). This leads to predictions from individual trees that are less correlated and, as a result, to a greater decrease in the variance of the estimates of the forest as a whole.

Similar to random forests, gradient boosting machines fit trees to subsets of the data points and subsets of the predictors. Whereas trees in random forests are grown independently, however, trees in gradient boosting machines are grown sequentially, based on the shortcomings of the previous trees. The first tree is fitted to the dependent variable, as is the case in random forests. However, the second tree is grown on the residuals of the first tree. After fitting the second tree, the predictions of the model are updated and a third tree is fitted on the residuals of these predictions. This process continually reduces the residuals of the model and is referred to as boosting.

When using the square loss function, the residuals of the model are equivalent to the negative gradient of the loss function. Each tree tries to minimize the residuals and consequently to maximize the gradient of the loss function. Fitting trees to the residuals thus incorporates the idea of gradient descent in a boosting algorithm. For other loss functions than the square loss function, the negative gradient is not equal to the residuals. In this case, fitting trees to the negative gradient of the loss function rather than the residuals is typically preferred, because the negative gradient tends to be less sensitive to outliers than the residuals. Given the use of gradient descent to minimize the loss function, the type of boosting described here is commonly known as gradient boosting.

For the analyses reported below, we use the implementation of gradient boosting machines (hereafter GBMs) in version 0.4−3 of the `xgboost` package for the statistical software R (T. Chen et al., 2015). The `xgboost` package allows for missing data and thus for a comprehensive analysis of a dependent variable that includes both one-character words and two-character words, despite the fact that predictors describing lexical properties of the second character are not available for one-character words.

The output of GBMs that we are interested in for the current purposes is the relative importance of predictors. The importance of a variable x is assessed through the average information gain (i.e., the decrease in entropy for the dependent variable) for all splits on x in all trees. Variable importances in an `xgboost` model are quite robust to collinearity in the input data, which makes the `xgboost` algorithm

particularly well-suited for use with the many highly correlated predictors in the CLD. In the following, we report variable importances for individual predictors as percentages of the total information gain obtained by all variables in the model. The total variable importance for all predictors, thus, sums up to 100%.

We ran GBMs for naming latencies, pronunciation durations and eye fixation durations. All GBMs consisted of 500 trees and were run with the standard parameter settings. The implementation of GBMs in the `xgboost` package does not take categorical variables as input. As recommended by the author of the `xgboost` package, we therefore replaced factor levels with conditional dependent variable means. For instance, the average pronunciation duration for words in which the first character has tone 1 was 456.39 ms. Therefore, we replaced the value `Tone 1` for `Character 1 Tone` with 456.39 for the pronunciation duration model. The other values for `Character 1 Tone` – and all other categorical predictors – were replaced in a similar fashion.

To control for the effects of experimental factors, we included experimental session (`Session`) and trial number (`Trial`) as predictors in all GBM analyses. In addition, since acoustic properties of the initial and final phoneme are known to influence naming latency and pronunciation duration measurements (see e.g., Y. Liu et al., 2006, 2007; Y. N. Chang et al., 2016), the initial and final phoneme of the word were included as predictors in all analyses (`Initial Phoneme`, `Final Phoneme`). Finally, we included the horizontal (`X Position`) and vertical (`Y Position`) position of a fixation (measured in pixels from the left edge and the top of the screen), as well as the temporal onset of a fixation relative to the onset of the stimulus (`Fixation Start Time`) as control variables in the eye fixation duration analysis.

Prior to all `xgboost` analyses, we removed incorrect responses (5.63%) as well as naming latencies (1.02%) and pronunciation durations (0.81%) that were further than 3 standard deviations from the naming latency and pronunciation duration means from the data. This led to a total data loss of 7.46% for the naming latency and pronunciation duration analyses.

For the eye fixation duration analysis, the removal of fixations with incorrect responses and naming latencies and pronunciation duration outliers resulted in the exclusion of 8.11% of the fixations. For this analysis, we furthermore removed fixations that were outside the region of the screen in which characters were displayed (0.40%). Finally, fixation duration outliers further than 3 standard deviations from the fixation duration mean were removed from the data (0.33%). Therefore, for the

eye fixation analysis, the total data loss for the eye fixation duration analysis was 8.84%.

3.3.2 Generalized additive models

GBMs provide insight into the quantitative contribution of lexical predictors to a dependent variable of interest. This provides valuable information about the relative importance of lexical predictors for the different measures of lexical processing. However, GBMs are less well-suited for investigating the qualitative nature of the effects of different lexical predictors.

In addition to the GBM analyses described above, we also carried out regression analyses through the use of generalized additive models (GAMs; Hastie & Tibshirani, 1986; S. Wood, 2006; S. N. Wood, 2011). GAMs are an extension of generalized linear models and allow for non-linear predictor effects without any predefined structure. Furthermore, GAMs offer the opportunity to model non-linear interactions between numerical predictors through the use of tensor products. As a result, GAMs provide a more detailed picture of the qualitative nature of predictor effects as compared to linear regression models or regression models that account for non-linearities through predefined structures (e.g., n-th order polynomials).

In the context of the current data set, regression models have two important shortcomings. First, regression models are not able to handle missing data. Typically, observations with missing values for one or more predictors (or dependent variables) are either omitted from the data or imputed. The problem of missing data is particularly relevant in the context of the current data due to the fact that predictors related to lexical properties of the second character are by definition missing for one-character words.

Omitting observations with one or more missing values for the current data set would result in the removal of all one-character words. Imputation of lexical properties of the second character for one-character words is not an option either. Over 15% of the words in the CLD are one-character words. Hence, data imputation would result in imputation of over 15% of the values for a large number of predictors. From a statistical point of view, this is far from optimal. However, even more importantly, the imputation of lexical properties of the second character for one-character words would not make conceptual sense. To overcome the problem of missing data for one-character words, we therefore carry out separate regression analyses for one-character words and two-character words.

The decision to use separate analyses for one-character words and two-character words solves part of the missing data problem for the current data set. Nonetheless, the subsets of the data for one-character words and two-character words still contain a certain proportion of missing data. For both the first and the second character, the CLD contains 12 measures that describe lexical properties of the phonetic radical (11 numerical variables, 1 categorical variable). No less than 29.36% of the 4,710 one-character words do not have a phonetic radical. For the 25,935 two-character words, the situation is even more problematic, with 50.49% of the first characters and 51.03% of the second characters having no phonetic radical. As a result, data imputation for phonetic radical measures is not feasible.

A pre-analysis of the data indicated that the predictive power of lexical predictors describing properties of phonetic radicals was limited in most models. Thus, we excluded phonetic radical measures from the analyses reported here. We did, however, carry out similar post-hoc analyses for the subsets of the data for which phonetic radicals were available. Whenever an effect of a phonetic radical measures was significant in such a post-hoc analysis, we report this effect directly after the corresponding main analysis.

After removing phonetic radical measures from the data, 55 numerical variables remained for the one-character words, whereas 118 numerical variables remained for two-character words. The number of missing data points for these variables was limited. Overall, 0.86% and 0.17% of all data points were missing for one-character words and two-character words, respectively. For one-character words, most values were missing for `Frequency` and `CD`, with 8.41% of the data missing for both predictors. For two-character words, most data points were missing for measures describing the diphone frequency of the second character (`Character 2 Mean Diphone Frequency`, `Character 2 Min Diphone Frequency`, `Character 2 Max Diphone Frequency`, `Character 2 Initial Diphone Frequency`). For each of these measures, 2.30% of all data points were missing.

To prevent data loss, we imputed missing values for the numerical predictors that remained after removing phonetic radical measures from the data through k-nearest neighbours imputation as implemented in the `knnImputation` function of version 0.4.1 of the `DMwR` package for R (Torgo, 2010). The `knnImputation` function was used with the default parameters settings ($k = 10, \ldots$).

The second shortcoming of regression models in the context of the current data is collinearity. As discussed in Chapter 2, many of the pairwise correlations between predictors are high. Entering highly correlated predictors into a regression model simultaneously can lead to misinformed conclusions about the qualitative and quantitative nature of effects of individual predictors (see, e.g., L. Friedman & Wall, 2005; Wurm & Fisicaro, 2014).

To overcome the problem of multicollinearity in the current regression analyses we subjected the numerical predictors in the CLD to a principal components analysis. A data set describes an n-dimensional space, with n being the number of variables in the data set. Each predictor describes a dimension in this n-dimensional space. A principal components analysis projects the data set onto a lower dimensional space, with dimensionality k (where k is a parameter set by the user). Each dimension in this lower dimensional space is referred to as a principal component.

The principal components algorithm transforms the data from an n-dimensional to a k-dimensional space in a manner that preserves as much of the information in the original data as possible. This is achieved by basing the dimensions in the new, k-dimensional space, on the amount of variance explained by a new dimension. The first principal component (i.e., the first dimension of the new k-dimensional space) is defined as the dimension that captures most variance in the data set. Additional principal components similarly are defined to capture the maximum amount of variance in the original data.

Importantly, however, a restriction applies in the definition of principal component 2 to k. Principal component i (with i ranging from 2 to k) is defined as the dimension that captures most of the variance in the original data *and that is orthogonal to principal components* 1 *to* $i-1$. The consequence of this restriction is that the principal components of a data set are uncorrelated and can therefore be used as input to a regression analysis without having to worry about the issue of multicollinearity.

Underlyingly, a principal components analysis is based on the Pearson correlation matrix for the input data. Given that Pearson correlations work best with symmetrical distributions, we applied power transforms to each numerical predictor prior to the principal components analyses. Power transforms help normalize the distribution of a numerical variable. The most well-known power transform is the Box-Cox transform (Box & Cox, 1964). The Box-Cox transform, however, is defined only for predictors with strictly non-negative values. The range for many of

the variables in the CLD contains non-positive values. Therefore, we use an alternative power transform that allows for negative predictor values: the Yeo-Johnson power transform (Yeo & Johnson, 2000).

The Yeo-Johnson power transforms for the numerical predictors in the CLD were done through the `yjPower` function in version 2.1-2 of the `car` package for R (Fox & Weisberg, 2011). Separate, independent power transforms were carried out for each predictor. Optimal values for the transformation parameter λ were determined through the `powerTransform` function in the `car` package. Furthermore, we scaled the resulting power-transformed variables prior to the principal components analysis to prevent the choice of principal components from being dominated by predictors with larger ranges.

Conceptually, principal components are a weighted sum of observed variables. Each variable in the n-dimensional input data contributes to a different degree to the values of a principal component in k-dimensional space. The contribution of a variable to the values of a principal component i is referred to as the loading of that variable on principal component i and is defined on a scale of -1 (perfect negative correlation between the values of a predictor and the values of a principal component) to 1 (perfect positive correlation between the values of a predictor and the values of a principal component). In a standard principal components analysis, there are no constraints that guide the loading of variables on principal components: the loadings arise naturally from the projection of the input data onto a lower-dimensional space that preserves the maximum amount of variance in the n-dimensional input space.

The unstructured manner in which loadings arise in a basic principal components analysis can lead to problems with respect to the interpretability of principal components. Oftentimes, loadings are spread across the -1 to 1 range, which leads to principal components that partially cover the information encoded in many different variables in the input data. Indeed, standard principal components analyses on the numerical variables for one-character words and two-character words led to highly uninterpretable sets of principal components.

To obtain components that are better interpretable, a rotation can be applied to the principal components produced by a standard principal components analysis. We used the varimax rotation to obtain better interpretable components. The varimax rotation rotates the principal components in such a manner that each component has a small number of large loadings and a large number of small (near-zero) loadings. As a result, each component is associated with a limited number of vari-

ables in the input data. Importantly, the varimax rotation is an orthogonal rotation, which means that the components remain uncorrelated after the rotation. Despite the fact that these components are no longer principal components in a technical sense, we will use the term principal component, or PC in short, to refer to the rotated components below.

The principal components analyses with varimax rotation were carried out using the `principal` function in version $1.5-8$ of the `psych` package for R (Revelle, 2015). For one-character words, we limited the number of components to 50, whereas for two-character words we extracted 100 components. However, we did not enter all principal components into the regression analyses. Instead, we restricted the set of principal components to include only those principal components for which at least one variable in the input data had a loading of 0.60 or higher. This resulted in a set of 21 principal components for one-character words and 54 principal components for two-character words.

The 21 principal components for one-character words and the 54 principal components for two-character words were entered into the regression analyses along with 3 (`Character 1 Tone`, `Character 1 Type`, `Character 1 Structure`) and 6 (`Character 1 Tone`, `Character 1 Type`, `Character 1 Structure`, `Character 2 Tone`, `Character 2 Type`, `Character 2 Structure`) categorical variables. This led to a total of 24 predictors for the analyses for one-character words and 60 predictors for the analyses for two-character words. To be conservative, we used an α level of 0.001 in all regression analyses. This α level was Bonferroni-corrected to $\frac{0.001}{24} = 0.000042$ for the analyses for one-character words and $\frac{0.001}{60} = 0.000017$ for the analyses for two-character words. Predictors were included as model terms in the reported regression models when their effect was significant at these corrected α levels only.

The naming latency and pronunciation duration analyses are uni-dimensional measures of human behaviour that provide little information about the point in time at which effects are present. Eye fixations, by contrast, were measured throughout each trial, starting 500 ms before the visual onset of the word and ending 2000 ms after the onset of the word. This allows us to investigate which lexical predictors influenced lexical processing at different points in time.

The time course of lexical predictor effects in the eye-tracking signal is often investigated by looking at the duration of the first fixation after stimulus onset, the second fixation after stimulus onset, et cetera. Given that the time between

the presentation of the pre-trial fixation mark and the word was constant (500 ms), however, the participant often anticipated the presentation of the stimulus by starting first fixations prior to the onset of the word. As a result, substantial processing takes place during fixations that start before the onset of the stimulus. Consequently, an analysis as described above is less than optimal for the current data set.

Rather than basing our analyses on fixation indexes, we instead grouped fixations based on the point in time at which they started relative to stimulus onset. We divided the eye fixations into fixation start time windows of 200 ms, starting with analyses of fixations that started in the -400 to -200 ms time window (and that ended after stimulus onset) and ending with analyses of fixations that started in the 1400 to 1600 ms time window. We chose 200 ms time windows to maintain statistical power while ensuring that fixations were not spread out in time too much. The average number of fixations per time window was $10,393.50$ for two-character words and $1,448.20$ for one-character words.

All regression analyses were carried out using the implementation of generalized additive models (hereafter GAMs; Hastie & Tibshirani, 1986) provided by the R package `mgcv` (S. Wood, 2006; S. N. Wood, 2011). `Initial Phoneme` and `Final Phoneme` were included as random effect terms. The effects of categorical variables were modelled through parametric terms, whereas the effects of numerical variables were modeled through smooth terms. Tensor product interactions were modeled as partial effects through the use of the `ti()` function in the `mgcv` package. Whenever such interactions terms were included in a model, we also included main effect smooths for the interacting predictors in the model. To prevent uninterpretable results, we restricted all predictor smooths to 4th order non-linearities ($k = 4$). Similarly, tensor product interactions were restricted to 4th order non-linearities in both dimensions.

Similar to the analyses with gradient boosting machines, we included experimental session (`Session`), trial number (`Trial`) and the initial (`Initial Phoneme`) and final phoneme (`Final Phoneme`) as control variables in all GAMs. The effects of experimental session and trial number were included as main effect smooths, whereas the initial and final phoneme were included as random intercepts. Furthermore, we included the horizontal (`X Position`) and vertical (`Y Position`) position of a fixation as control variables in the eye fixation duration analyses.

Incorrect naming responses were removed from the data prior to analysis. This led to the exclusion of 14.69% of the data for one-character words and 3.99% of the data for two-character words, for an overall error rate of 5.63%. In addition, reaction time and pronunciation duration outliers that were further than 3 standard deviations from the reaction time and pronunciation duration means for correct responses were removed prior to all analyses. Reaction time and pronunciation duration outliers were removed separately for one-character words and two-character words. This resulted in the exclusion of a further 0.47% for one-character words and 0.98% of the data for two-character words with extreme naming latencies (overall rejection rate naming latencies: 0.90%) and an exclusion of a further 0.55% of the data for one-character words and 0.75% of the data for two-character words with extreme pronunciation durations (overall rejection rate pronunciation durations: 0.72%). In total, this led to a data loss of 15.71% for one-character words and 5.71% for two-character words, for an overall data loss of 7.25%.

For the fixation duration analysis, we removed fixations that were outside the region of the screen in which characters were displayed. For one-character words, this led to the exclusion of 1.26% of all fixations, whereas for two-character words 0.28% of all fixations were rejected. Finally, we removed fixation duration outliers further than 3 standard deviation from the fixation duration mean. Again, we separately removed outliers for one-character words and two-character words. As a result, 0.08% of all fixations were removed for one-character words, whereas 0.35% of all fixations were removed for two-character words. Overall, 0.41% of all fixations were removed for being outside the region of the screen in which characters were displayed, whereas 0.31% of all fixations were rejected as fixation duration outliers. In total, the data loss for the eye fixation duration analyses was 19.62% for one-character words and 6.99% for two-character words, for an overall data loss of 8.73%.

We did not remove predictor outliers prior to analysis to prevent further data loss. However, for all reported models, we fitted an identical model to the subset of the data for which the residuals of the original model were within 2.5 standard deviations of the mean of the residuals for the model in question. This had an influence on the significance of a single model term across all reported models only. For this model term, we explicitly mention the lack of significance when residual outliers are removed from the data. Model summaries for all reported models are presented in Appendix A.

3.4 Results

In the following, we describe the results for the word naming experiment. We first describe the analyses for the naming latencies, followed by the analyses for the pronunciation durations and eye fixation durations. For each of these 3 behavioural measures, we first present the results of the gradient boosting machine (GBM) analysis. Next, we present the results of the analyses using generalized additive models (GAMs) for both one-character words and two-character words.

3.4.1 Naming latencies

3.4.1.1 GBM

Table 3.1 presents the relative influences for the 153 predictors entered into the gradient boosting machine on the naming latencies. As mentioned in the Analysis section, variable importances were converted to percentages, such that the total variable importance for all predictors is 100%. We provide the full table of relative influences for all predictors, but limit our discussion of the results to the 20 predictors with the highest relative influence.

As can be seen in Table 3.1, the identity of the initial phoneme was the strongest predictor for naming latencies, with a relative influence of 15.441%. This highlights the importance of controlling for differences in naming latencies due to the acoustic properties of word pronunciations. The experimental control variables Session (8.637%) and Trial (1.646%) similarly account for a significant portion of the variance in the naming latencies.

The importance of the control variables is reflected in Figure 3.1, which shows the summed relative influence per cluster (outer circle) and group (inner circle) of predictors (see Chapter 2 for a detailed description of these groups and clusters). The control variables are categorized as "other" in this figure and have a summed relative influence of 25.759%.

Figure 3.1 furthermore indicates that frequency measures (displayed in blue) are the strongest type of predictor by a large margin. The 6 clusters with frequency measures have a summed relative influence of no less than 54.982%. Most of this predictive power comes from the lexical variables in Cluster 3, which contains measures of the frequency of the first character (summed relative influence: 32.989%). Correspondingly, the contextual diversity of the first character in the SCCoW (14.395%) and SUBTLEX-CH (5.834%) corpora are the strongest lexical predictors. Further-

Table 3.1: Relative variable influences in an XGBoost model fitted to the naming latencies. Abbreviations: BE = Backward Enemies, C1 = Character 1, C2 = Character 2, Diph. = Diphone, Freq. = Frequency, HLC = High-Level Components, LLC = Low-Level Components, Phono = Phonological, Phon. = Phoneme, SUBTL = SUBTLEX-CH, Typ. = Types, Tok. = Tokens.

rank	predictor	infl.	rank	predictor	infl.
1	Initial Phoneme	15.441	29	C2 Freq. (Gigaword)	0.493
2	C1 CD (SCCoW)	14.395	30	C1 Family Freq.	0.492
3	Session	8.637	31	C2 Structure	0.408
4	C1 CD (SUBTL)	5.834	32	C2 CD (SCCoW)	0.399
5	CD (SUBTL)	4.127	33	C2 CD (Gigaword)	0.370
6	C1 Freq. (SUBTL)	3.906	34	PMI	0.370
7	C1 Friends	3.485	35	C1 Picture Size	0.370
8	C1 Strokes	2.514	36	C1 PR Enemies Freq.	0.358
9	C2 Freq. (SUBTL)	2.171	37	C2 Pixels	0.350
10	C1 CD (Gigaword)	1.992	38	C1 Pixels OLD	0.344
11	Frequency (SCCoW)	1.747	39	C1 Freq. (SCCoW)	0.327
12	C1 Entropy	1.711	40	C1 RE	0.319
13	Trial	1.646	41	C1 Family Size	0.319
14	C1 Freq. (Gigaword)	1.491	42	C1 PR Frequency	0.318
15	Strokes	1.367	43	t-Score	0.306
16	CD (SCCoW)	1.289	44	Entropy Char. Freqs.	0.299
17	Frequency (Gigaword)	1.274	45	CD (Gigaword)	0.287
18	C2 Family Size	1.189	46	C1 Mean HLC Freq.	0.286
19	C2 Entropy	1.103	47	C1 SR Frequency	0.275
20	Frequency (SUBTL)	0.994	48	C2 Pixels OLD	0.274
21	C1 Trigram Entropy	0.876	49	C1 Homophones Freq.	0.259
22	C1 LLC	0.862	50	C1 Friends Freq.	0.258
23	C2 Friends	0.825	51	C2 Family Freq.	0.257
24	C2 Freq. (SCCoW)	0.782	52	C1 Initial Phon. Freq.	0.255
25	Position-specific PMI	0.756	53	C1 Homographs Freq.	0.253
26	C1 Pixels	0.705	54	C1 PR Friends	0.250
27	C1 Traditional Freq.	0.543	55	C1 PR Friends Freq.	0.240
28	C1 LLC OLD	0.518	56	C1 Mean Phon. Freq.	0.235

Table 3.1 (continued)

rank	predictor	infl.	rank	predictor	infl
57	Traditional Freq.	0.233	90	C2 PR Frequency	0.110
58	C1 Min LLC Freq.	0.229	91	C1 HLC	0.106
59	Mean Phon. Freq.	0.220	92	C2 Friends Freq.	0.104
60	C1 Structure	0.217	93	C1 LLC N	0.103
61	C2 Trigram Entropy	0.217	94	PLD	0.102
62	C1 SR Family Size	0.204	95	C1 PLD	0.102
63	Mean Diph. Freq.	0.202	96	C2 Mean HLC Freq.	0.100
64	C1 Phono N	0.202	97	C1 PR Enemies (Tok.)	0.100
65	C1 Mean LLC Freq.	0.202	98	C2 PR Enemies (Tok.)	0.097
66	C2 RE	0.201	99	C2 PR BE Freq.	0.096
67	C2 Mean Diph. Freq.	0.184	100	C2 PR Family Size	0.092
68	Max Diph. Freq.	0.183	101	C2 Min LLC Freq.	0.090
69	C1 Min HLC Freq.	0.178	102	C1 Initial Diph. Freq.	0.088
70	Min Diph. Freq.	0.167	103	C2 HLC	0.085
71	C2 SR Frequency	0.160	104	C2 Max HLC Freq.	0.084
72	C2 Min HLC Freq.	0.156	105	C2 PR BE (Tok.)	0.082
73	C1 PR BE Freq.	0.153	106	C1 Homographs (Typ.)	0.077
74	C2 Homophones Freq.	0.151	107	C1 Max Phon. Freq.	0.075
75	Transitional Diph. Freq.	0.147	108	C2 LLC OLD	0.073
76	C2 Phono N	0.145	109	C1 Homophones (Tok.)	0.072
77	C1 Mean Diph. Freq.	0.142	110	C2 Homographs Freq.	0.070
78	C2 Traditional Freq.	0.141	111	C2 SR Strokes	0.069
79	C2 Min Diph. Freq.	0.137	112	C1 PR Enemies (Typ.)	0.067
80	C2 Mean LLC Freq.	0.133	113	C1 Max HLC Freq.	0.064
81	C2 PR Friends Freq.	0.132	114	C1 Homophones (Typ.)	0.061
82	C1 Min Diph. Freq.	0.131	115	C2 CD (SUBTL)	0.059
83	C2 Picture Size	0.130	116	Phono N	0.057
84	C1 PR Strokes	0.129	117	Min Phon. Freq.	0.056
85	C1 PR BE (Tok.)	0.125	118	C1 Min Phon. Freq.	0.056
86	C1 PR Family Size	0.122	119	C2 PR Enemies Freq.	0.055
87	C2 Mean Phon. Freq.	0.121	120	C2 Strokes	0.054
88	C2 SR Family Size	0.119	121	C2 Homophones (Tok.)	0.054
89	C1 Homograph (Tok.)	0.115	122	C2 Initial Phon. Freq.	0.052

Table 3.1 (continued)

rank	predictor	infl.	rank	predictor	infl.
123	C1 Max Diph. Freq.	0.049	139	C1 Max LLC Freq.	0.026
124	C1 Tone	0.046	140	C2 Max LLC Freq.	0.025
125	C2 PR Friends	0.045	141	C1 PR Regularity	0.020
126	C1 Phonemes	0.044	142	C2 Min Phon. Freq.	0.019
127	C2 Max Diph. Freq.	0.042	143	C2 PLD	0.018
128	C2 Type	0.041	144	C2 LLC	0.017
129	C1 SR Strokes	0.038	145	C2 PR Enemies (Typ.)	0.016
130	C2 PR Strokes	0.037	146	C2 Initial Diph. Freq.	0.015
131	C1 PR BE (Typ.)	0.036	147	C2 Homographs (Typ.)	0.014
132	C2 Homograph (Tok.)	0.036	148	C2 LLC N	0.013
133	C2 Homophones (Typ.)	0.035	149	C2 PR BE (Typ.)	0.012
134	Final Phoneme	0.035	150	C1 Type	0.007
135	Phonemes	0.034	151	C2 Phonemes	0.006
136	C2 Tone	0.029	152	Length	0.000
137	C2 Max Phon. Freq.	0.028	153	C2 PR Regularity	0.000
138	Max Phon. Freq.	0.027			

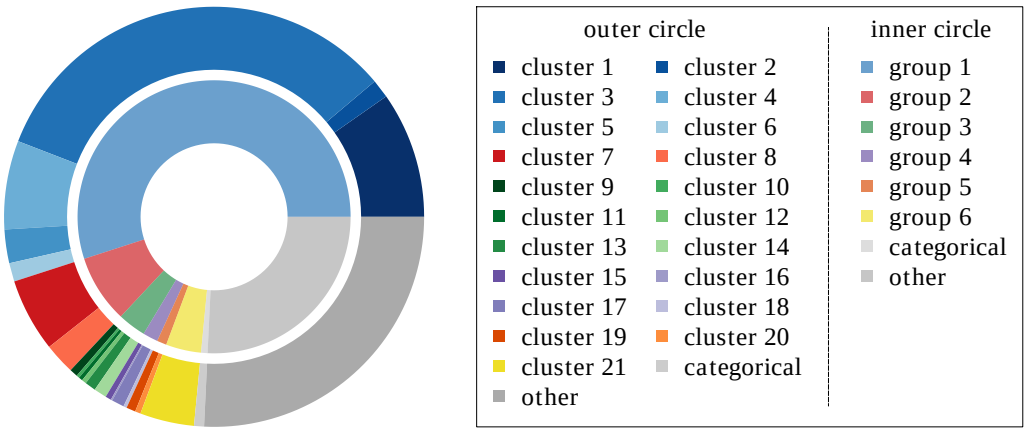

Figure 3.1: Summed relative influence per cluster (outer circle) and per group of clusters (inner circle) of predictors in an XGBoost model fitted to the naming latencies.

more, the frequency of the first character in the SUBTLEX-CH corpus (3.906%), the number of friends of the first character (3.485%), the contextual diversity of the first character in the Gigaword corpus (1.992%) and the frequency of the first character in the Gigaword corpus (1.491%) are among the twenty predictors with the highest relative influence.

Frequency measures of the word as a whole (Cluster 1; 9.718%) and of the second character (Cluster 4; 6.826%) contribute to a lesser, but nonetheless considerable degree to the explanatory power of the gradient boosting machine. The frequency of the word as a whole in the SCCoW (1.747%), the Gigaword corpus (1.274%) and the SUBTLEX-CH (0.994%) corpus are all among the twenty strongest predictors, as are the contextual diversity of the word as a whole in the SUBTLEX-CH corpus (4.127%) and in SCCoW (1.289%). The strongest predictors for the frequency of the second character are the frequency in the SUBTLEX-CH corpus (2.171%) and the character's family size (1.189%).

Out of the 20 predictors with the highest relative influence, no less than 12 measures are contextual diversity and frequency measures. Six of these measures were derived from the SUBTLEX-CH corpus, 3 from the SCCoW and 3 from the Gigaword corpus. Therefore, the SUBTLEX-CH frequency measures contribute strongly to the explanatory power of the GBM. However, it should be noted that the summed relative influence for the frequency and contextual diversity measures is greater for SCCoW (18.939%) than for SUBTLEX-CH (17.091%). This is the case despite the fact that the SCCoW measures correlate more strongly with the Gigaword measures (summed relative influence: 5.907%) than do the SUBTLEX-CH measures and therefore are in fiercer competition with these measures when split decisions are made in the trees of the GBM.

Interestingly, the summed relative influence of the contextual diversity measures from the three corpora (28.752%) is much greater than the summed relative influence of the frequency measures (13.185%). That is, contextual diversity is a stronger predictor of naming latencies as compared to frequency. This finding is in line with the findings of Cai and Brysbaert (2010) for word naming latencies for a set of $2,289$ single character words and with the correlations of the frequency and contextual diversity measures with the naming latencies (see Table 2.6). However, the correlations for the current data as well as for the data discussed in Cai and Brysbaert (2010) only showed subtle differences between frequency and contextual diversity measures. By contrast, the analysis using a GBM indicates that when these

measures compete for inclusion in a tree-based model the difference is substantial, with a much larger contribution of contextual diversity measures than of frequency measures.

The influence of the other three clusters in the group of frequency-related clusters is more limited, with a summed relative influence of 1.432% for the association measures in Cluster 2 and summed relative influences of 2.587% and 1.430% for the clusters containing entropy measures of the first and second character, respectively. The influence of the measures in Cluster 5 and Cluster 6 is mostly driven by the entropy of the first character (1.711%) and the second character (1.103%). In other words: the predictability of the second character given the first character, and vice versa influences naming latencies to a considerable degree.

Figure 3.1 indicates that the influence of the other 5 groups of clusters is much more limited than that of the group of clusters containing frequency-based measures. The only other group of clusters for which individual measures are among the 20 predictors with the highest relative influence is Group 2, which contains clusters that describe the visual complexity of a word and its characters. The summed relative influence for this group of clusters is 8.051%.

The summed relative influence of the measures in Cluster 7 (visual complexity of the first character) is 5.651%, with the number of strokes in the first character being the strongest predictor in this cluster (relative influence: 2.514%). The summed relative influence of the measures in the other cluster in Group 2, Cluster 8 (visual complexity of the second character and the word as a whole), is 2.4%. The strongest predictor in this cluster is the number of strokes in the word as a whole, with a relative influence of 1.367%.

The fact that the number of strokes is most predictive for naming latencies indicates that the stroke level may be the optimal level for measuring the effects of visual complexity as far as the lexical processes that influence naming latencies are concerned. The summed relative influence of the number of strokes of the first and second character and the word as a whole is 3.935%. At a higher level, the summed relative influence of the number of high-level components (summed relative influence: 0.191%) and the number of low-level components (summed relative influence: 0.879%) is much more limited. At a lower level, the summed relative influence of the number of pixels of the first and second character (1.055%) and the file size of image files displaying the first and second character (0.500%) are more modest as well. Stroke count thus seems to be a satisfactory measure of visual complexity that is not easily improved upon by defining visual complexity at a different grain size.

The lexical variables in the group of clusters related to the phonology of a word and its characters (Group 3, summed relative influence: 3.341%), as well as the lexical variables in the groups of clusters that describe the number of homographs (Group 4, summed relative influence: 1.79%) and homophones (Group 5: summed relative influence: 1.136%) contribute little to the explanatory power of the GBM and the same holds for the lexical variables in the cluster of "other" numerical predictors in Group 6 (summed relative influence: 4.174%). The summed relative influence of the categorical predictors that were not included in the clustering technique applied to the SOM described in Chapter 2 was likewise limited (summed relative influence: 0.768%).

In conclusion, the GBM analysis suggests that naming latencies are co-determined primarily by frequency-based measures. The frequency of the first character is pivotal, but the frequency of the word as a whole and the second character play an important role as well. Furthermore, the predictability of one character given the other character also co-determines naming latencies. In addition to frequency-based measures, the visual complexity of the input explains part of the variance in the naming latencies, particularly at the stroke level. However, the GBM analysis says little about the qualitative nature of the effects of these measures. Therefore, we now turn to the analysis of the naming latencies with GAMs.

3.4.1.2 GAMs

3.4.1.2.1 One-character words

The GAM model fitted to the naming latencies for one-character words showed significant effects of the control variable `Session` ($F = 64.779$, $p < 0.001$) and `Initial Phoneme` ($F = 14.208$, $p < 0.001$). Naming latencies were shorter for words that started fricatives and longer for words that started with plosives or vowels. The effect of `Session` is presented in Figure 3.2. Although the model was fitted to inverse-transformed latencies, we show the effect – and all effects described below – on the original predictor scale for ease of interpretation. At the start of the experiment naming latencies are relatively long. The average naming latency steadily decreases over the first ten sessions of the experiment. After the tenth session the naming latencies stabilize, which suggests that the participant reached optimal performance after naming about $10,000$ words. The effect size for the effect of `Session`, defined as the difference between the highest and lowest predicted value, is 67 ms.

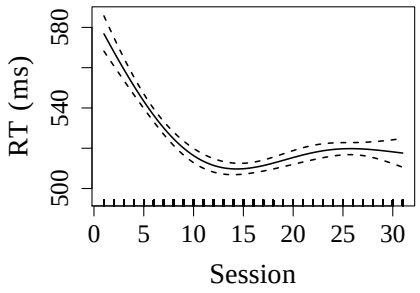

Figure 3.2: Reaction time results: one-character words. Experimental predictors.

In addition to the effects of these control variables, the GAM showed significant effects of three principal components at the Bonferroni-corrected α level. Consistent with the results from the GBM analysis, the principal component with the strongest effect on the naming latencies is `PC1` ($F = 206.977$, $p < 0.001$), which describes the frequency of the first (and only) character and the frequency of the word.

The predictors with the highest loadings on `PC1` are `Character 1 CD (SCCoW)` (0.972), `Character 1 Frequency (SCCoW)` (0.966), `Character 1 CD (Gigaword)` (0.962), `Character 1 Frequency (Gigaword)` (0.959), and `Frequency (Gigaword)` (0.945). The contextual diversity and frequency in the SUBTLEX-CH also have high loadings on PC1 (`Character 1 CD (SUBTLEX-CH)`: 0.927; `Character 1 Frequency (SUBTLEX-CH)`: 0.919), as do the contextual diversity and frequency of the word in the SCCoW and in SUBTLEX-CH (`CD (SCCoW)`: 0.842; `Frequency (SCCoW)`: 0.866; `CD (SUBTLEX-CH)`: 0.828; `Frequency (SUBTLEX-CH)`: 0.822). Therefore, `PC1` provides a composite measure of the contextual diversity and frequency measures from the SCCoW, the Gigaword corpus and SUBTLEX-CH.

The effect of `PC1` is presented in Figure 3.3. Consistent with previous findings (see e.g., Y. N. Chang et al., 2016; Cai & Brysbaert, 2010; Y. Liu et al., 2007), a higher frequency leads to shorter naming latencies. The effect is most pronounced for low predictor values and levels off for high frequencies. The effect size is large, with a difference of 190 ms between the predicted naming latencies for low and high values of `PC1`.

In addition to `PC1`, we also observed an effect of `PC2` ($F = 37.262$, $p < 0.001$). `PC2` describes the visual complexity of the first character, with high loadings for the number of low-level and high-level components (`Character 1 Low-Level Components`: 0.901, `Character 1 High-Level Components`: 0.782), the number of strokes

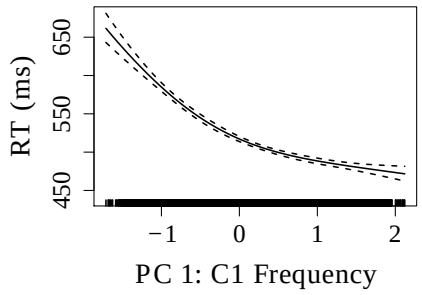

Figure 3.3: Reaction time results: one-character words. Character 1 frequency.

(`Character 1 Strokes`: 0.849) and the number of pixels (`Character 1 Pixels`: 0.767) of the first character, as well as for the orthographic Levenshtein distance at the level of the low-level components (`Character 1 Low-Level Components OLD`: 0.865). Figure 3.4 shows the effect of `PC2`. The effect is near-linear, with naming latencies being 78 ms longer for highly complex characters as compared to visually simple characters. Again, this pattern of results is comparable to the effects of visual complexity observed in previous studies (see e.g., Y. N. Chang et al., 2016; Y. Liu et al., 2007).

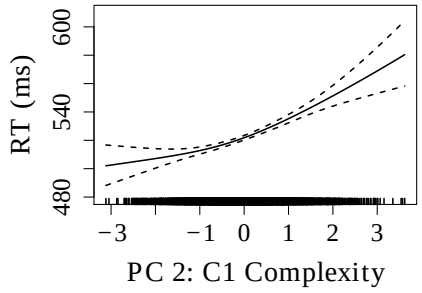

Figure 3.4: Reaction time results: one-character words. Character 1 complexity.

The third principal component that showed a significant effect on the naming latencies, `PC11`, describes the frequency in traditional Chinese, with high loadings for both `Character 1 Traditional Frequency` (0.925) and `Traditional Frequency` (0.896). Both of these predictors had moderate contributions to the GBM model, with relative influences of 0.493 and 0.235, respectively.

The effect of `PC11` ($F = 51.694$, $p < 0.001$) is shown in Figure 3.5. The effect is inverse U-shaped, with long naming latencies for words with medium character and word frequencies in traditional Chinese and shorter naming latencies for words

with both high and low character and word frequencies in traditional Chinese. The difference between the highest and lowest predicted values is 87 ms.

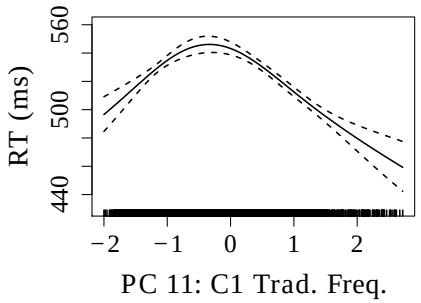

Figure 3.5: Reaction time results: one-character words. Character 1 traditional frequency.

The qualitative nature of the effect of the frequency in traditional Chinese is rather different from the typical facilitatory effect of frequency. The nature of the traditional frequency effect may be a result of the encoding of traditional frequency measures in the CLD. For words and characters that did not occur in the Academia Sinica Corpus, we set the corresponding traditional frequency counts to 0. More often than not, however, a traditional frequency count of 0 implies that the character or word was simplified and therefore does not exist in traditional Chinese.

The pattern of results for `PC11`, then, makes sense. For one-character words that do not exist in traditional Chinese, naming latencies are intermediate. For words that do exist in traditional Chinese, but that have relatively low frequencies, naming latencies are long. Finally, for words that exist in traditional Chinese and that have a high frequency, naming latencies are short. When taking one-character words that were simplified into account, the effect of traditional frequency thus shows the expected near-linear facilitation.

Finally, a post-hoc analysis on the subset of the one-character words that contain a phonetic radical revealed an effect of the number of friends of the phonetic radical and their frequency. The corresponding principal component, `PC13P`[1], has highest loadings for `Character 1 PR Friends` (0.812) and `Character 1 PR Friends Frequency` (0.748). The effect of `PC13P` ($F = 66.003$, $p < 0.001$) is presented in Figure 3.6.

[1]Principal components were recalculated for the subset of the data for which phonetic radical measures were available. To avoid confusion with the principal components in the main analysis we add a suffix `P` to the principal components from this post-hoc analysis.

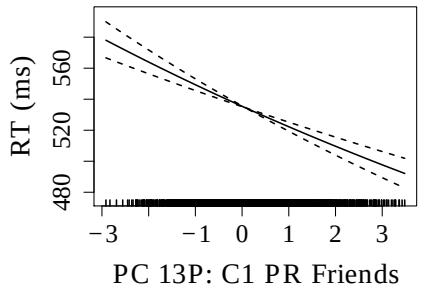

Figure 3.6: Reaction time results: one-character words (post-hoc analysis). Character 1 PR friends.

The effect of `PC13P` is linear in nature and has an effect size of 86 ms. Naming latencies are shorter for characters that share their pronunciation with a greater number of other characters that contain the same phonetic radical. This effect is line with the relatively strong effects of phonetic radical regularity observed in previous single character naming studies (see e.g., Y. Liu et al., 2007; Y. N. Chang et al., 2016). The cross-character reliability of the information provided by the phonetic radical thus co-determines naming latencies for single character words to a considerable degree.

3.4.1.2.2 Two-character words

Similar to the GAM model for one-character words, the GAM model fitted to the naming latencies for two-character showed effects of `Initial Phoneme` ($F = 115.306$, $p < 0.001$) and `Session` ($F = 156.845$, $p < 0.001$). As was the case for one-character words, naming latencies were shorter for fricatives and longer for vowels and plosives. The effect of `Session` is presented in the left panel of Figure 3.7. The effect is highly similar to the effect of `Session` for one-character words, with a facilitatory effect that levels off after 10 experimental sessions. However, at 32 ms the effect size of the effect of `Session` for two-character words is much smaller than the effect size of `Session` for one-character words (67 ms).

In addition to the effects of `Initial Phoneme` and `Session`, we observed an effect of a third control variable: `Trial` ($F = 27.023$, $p < 0.001$). As can be seen in the right panel of Figure 3.7, the effect size for the effect of `Trial` is limited, with naming latencies for words near the end of an experimental session being 11ms shorter than naming latencies for words that appear earlier in a session.

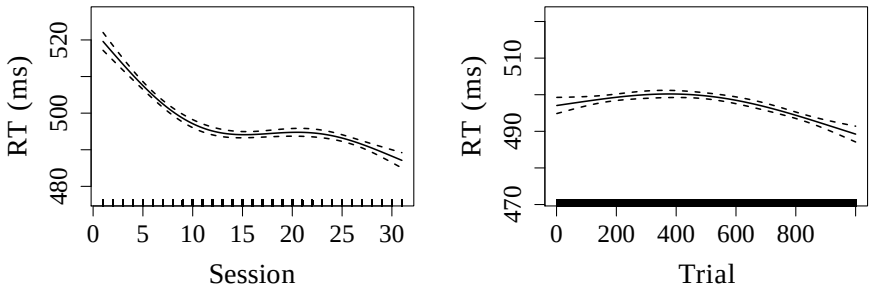

Figure 3.7: Reaction time results: two-character words. Session (left panel) and Trial (right panel).

The interaction between `Session` and `Trial` ($F = 6.292$, $p < 0.001$) sheds more light on the effect of `Trial`. Figure 3.8 presents the additive contour surface of the main effect for `Session`, the main effect for `Trial` and the interaction between `Session` and `Trial`. In the early experimental sessions, a clear effect of `Trial` is present, with naming latencies being up to 39 ms shorter for trials near the end of the experimental session. However, for later experimental sessions, no such effect is present. Therefore, the facilitatory effect of `Trial` is no longer present when the participant has fully adapted to the experimental paradigm.

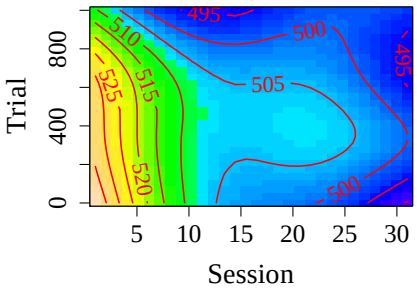

Figure 3.8: Reaction time results: two-character words. Interaction between Session and Trial. Additive contour surface of the main effect for Session, the main effect for Trial and the interaction between Session and Trial.

For one-character words, we found a strong effect of `PC1`, which encoded the overall frequency of the character, as well as the frequency of the character when it was used as an autonomous word. For two-character words, the frequency of the first and second character is encoded in `PC1` and `PC2`, respectively. The lexical predictors with the highest loading on `PC1` are the frequency and contextual diversity of the

first character in the SCCoW (`Character 1 Frequency (SCCoW)`: 0.959, `Character 1 CD (SCCoW)`: 0.958). However, the frequency and contextual diversity measures from the Gigaword corpus and SUBTLEX-CH are represented by `PC1` as well, with loadings of 0.949 and 0.878 for `Character 1 Frequency (Gigaword)` and `Character 1 Frequency (SUBTLEX-CH)` and loadings of 0.952 and 0.882 for `Character 1 CD (Gigaword)` and `Character 1 CD (SUBTLEX-CH)`. Similarly, the predictors with the highest loading on `PC2` are the frequency and contextual diversity of the second character in the SCCoW (`Character 2 Frequency (SCCoW)`: 0.959, `Character 2 CD (SCCoW)`: 0.958). Again, the corresponding measures from the Gigaword corpus and the SUBTLEX-CH corpus have high loadings on `PC2` as well (`Character 2 Frequency (Gigaword)`: 0.947; `Character 2 Frequency (SUBTLEX-CH)`: 0.859; `Character 2 CD (Gigaword)`: 0.950; `Character 2 CD (SUBTLEX-CH)`: 0.873). Hence, the principal components `PC1` and `PC2` provide a composite measures of the character 1 and character 2 frequency and contextual diversity measures in the SCCoW, the Gigaword corpus and SUBTLEX-CH.

For one-character words, we found a facilitatory effect of frequency that was most pronounced at the lower part of the frequency range and that levelled off for high values of frequency. For two-character words, the effects of character frequency are highly similar. Figure 3.9 presents the effects `PC1` (left panel) and `PC2` (right panel). The effects of both `PC1` ($F = 162.057$, $p < 0.001$) and `PC2` ($F = 237.189$, $p < 0.001$) are characterized by shorter naming latencies for higher predictor values, with the effect being most prominent in the lower part of the predictor range. Consistent with the GBM analysis, the effect size is somewhat larger for `PC1` (100 ms) than for `PC2` (79 ms). This suggests that naming latencies are determined to a greater extent by the frequency of the first character as compared to the frequency of the second character.

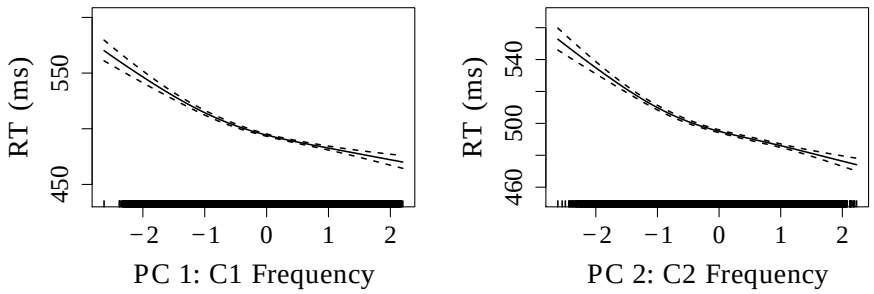

Figure 3.9: Reaction time results: two-character words. Character 1 frequency (left panel) and character 2 frequency (right panel).

For one-character words, the overall frequency of the character and the frequency of the character used as a word are highly correlated (correlation between `Character 1 Frequency (SCCoW)` and `Frequency (SCCoW)` for all one-character words in the CLD: $r = 0.812$). For two-character words, the correlation between the frequency of the word as a whole and the frequency of the characters is much lower (correlation between `Character 1 Frequency (SCCoW)` and `Frequency (SCCoW)`: $r = 0.138$; correlation between `Character 2 Frequency (SCCoW)` and `Frequency (SCCoW)`: $r = 0.146$). As a result, the frequency of the word as a whole was encoded in a separate principal component: `PC5`. Interestingly, this principal component has high loadings for the frequency and contextual diversity in the SCCoW and the Gigaword corpus (`Frequency (SCCoW)`: 0.941; `Frequency (Gigaword)`: 0.935; `CD (SCCoW)`: 0.945; `CD (Gigaword)`: 0.938), but medium loadings for the frequency and contextual diversity in SUBTLEX-CH (`Frequency (SUBTLEX-CH)`: 0.583; `CD (SUBTLEX-CH)`: 0.595) only. We return to this issue shortly.

The effect of `PC5` ($F = 200.210$, $p < 0.001$) is presented in the left panel of Figure 3.10. The effect of word frequency is qualitatively highly similar to that of character 1 and character 2 frequency and shows a facilitatory effect that levels off for high frequency words. The effect size of word frequency is somewhat smaller than that of the character frequencies, with a difference in predicted values for the lowest frequency words and the highest frequency words of 50 ms.

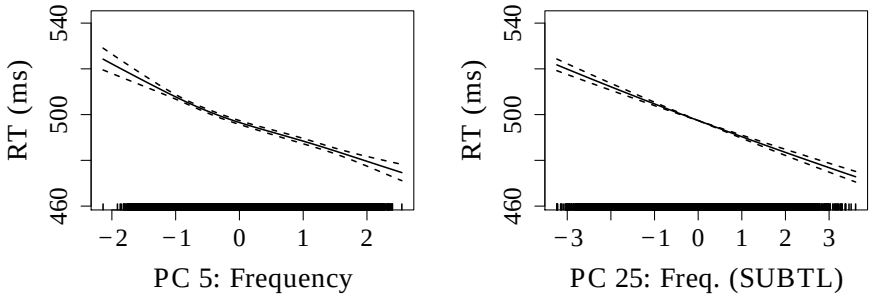

Figure 3.10: Reaction time results: two-character words. Word frequency (left panel) and word frequency in the SUBTLEX-CH corpus (right panel).

The right panel of Figure 3.10 shows the effect of another principal component that contains information about the frequency of the word as a whole: `PC25` ($F = 406.090$, $p < 0.001$). The frequency and contextual diversity in the SCCoW and the Gigaword have low loadings on `PC25` (`Frequency (SCCoW)`: 0.089; `Frequency`

(Gigaword): 0.087; CD (SCCoW): 0.082; CD (Gigaword): 0.078). By contrast, Frequency (SUBTLEX-CH) (0.796) and CD (SUBTLEX-CH) (0.785) have high loadings on PC25.

While frequency measures of each character are encoded in a single principal component, frequency measures of the word as a whole are split out over two principal components: one for the measures from SCCoW and the Gigaword corpus and another for the measures from SUBTLEX-CH. This implies that the character frequencies in the three corpora are highly similar, whereas the word frequencies measures from the SCCoW and the Gigaword corpus on the one hand and the word frequencies from SUBTLEX-CH on the other hand provide information that is at least partially different. At the character level, the SUBTLEX-CH frequencies indeed show strong correlations with the frequency measures from the SCCoW and the Gigaword corpus. For the first character, the correlations of the SUBTLEX-CH frequencies with the frequencies from the SCCoW and the Gigaword corpus are $r = 0.731$ and $r = 0.645$, respectively. For the second character, the equivalent correlations are $r = 0.701$ and $r = 0.647$. By contrast, the word-level SUBTLEX-CH frequencies show medium-strength correlations of $r = 0.396$ and $r = 0.335$ with the word frequencies from the SCCoW and the Gigaword corpus only. Therefore, the division of word-level frequency measures across two principal components follows straightforwardly from the differences in the distributional structure of frequency measures at the character level and the word level.

As can be seen in the right panel of Figure 3.10, the effect of PC25 is linear, with shorter naming latencies for words with higher frequencies in SUBTLEX-CH. With a 49 ms difference between the predicted values for the words with the lowest and highest frequencies in SUBTLEX-CH, the effect size of PC25 is highly similar to that of PC5 (50 ms). That is, the principal component based on the SUBTLEX-CH frequencies has similar explanatory power as compared to the word frequencies from the SCCoW and the Gigaword corpus.

In addition to the effect of frequency in simplified Chinese, the analysis for the one-character words showed an effect of the frequency in traditional Chinese. The analysis for the two-character words similarly showed an effect of traditional Chinese frequency, both at the character level and at the word level. Figure 3.11 presents the results of PC40 and PC28, which describe the first character frequency and the word frequency in traditional Chinese, respectively. PC40 has a high loading for Character 1 Traditional Frequency (0.843) and a low loading for Traditional Frequency

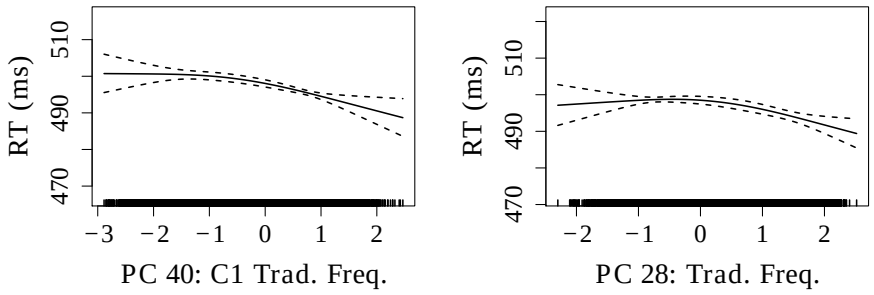

Figure 3.11: Reaction time results: two-character words. Traditional Chinese frequency for character 1 (left panel) and for the word as a whole (right panel).

(0.204). Conversely, `PC28` has a high loading for `Traditional Frequency` (0.927) and a low loading for `Character 1 Traditional Frequency` (0.287).

As can be seen in Figure 3.11, the effects of both `PC40` ($F = 18.857$, $p < 0.001$) and `PC28` ($F = 10.868$, $p < 0.001$) are facilitatory, with shorter naming latencies when the first character or word has a high frequency in traditional Chinese. The effects of `PC40` and `PC28` are most prominent for the upper half of the predictor ranges. For low predictor values, the effects levels off for both principal components. The effect sizes of the traditional frequency effects are limited. `PC40` has an effect size of 12 ms, whereas `PC28` shows a difference of a mere 9 ms between the highest and lowest predicted values.

In addition to frequency effects, the analysis for one-character words showed an effect of visual complexity, with longer naming latencies for more complex characters. We found similar effects for the complexity of the first and second character in two-character words, which is encoded in `PC4` and `PC3`. The predictors with the highest loadings on `PC3` are `Character 2 Low-Level Components` (0.874), `Character 2 Strokes` (0.864), `Character 2 Low-Level Components OLD` (0.817) and `Character 2 Pixels` (0.813). The predictors with the highest loadings on `PC4` are the first character counterparts of these measures: `Character 1 Low-Level Components` (0.874), `Character 1 Strokes` (0.837), `Character 1 Low-Level Components OLD` (0.838) and `Character 1 Pixels` (0.788). That is, `PC3` describes the visual complexity of the second character, whereas `PC4` encodes information about the visual complexity of the first character.

Figure 3.12 shows the effects of `PC4` (left panel; $F = 222.793$, $p < 0.001$) and `PC3` (right panel; $F = 58.266$, $p < 0.001$). The effects are qualitatively similar, with longer naming latencies for more complex characters. As was the case for the

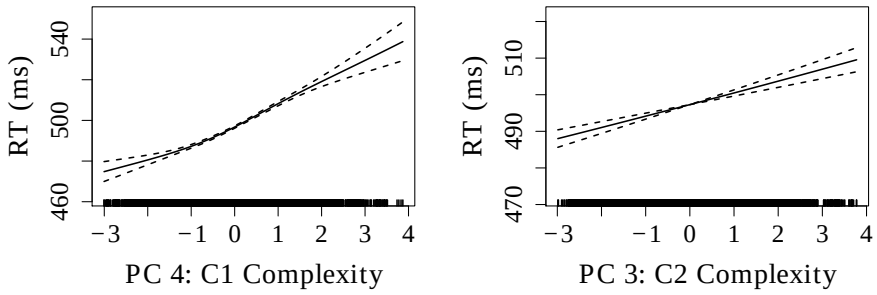

Figure 3.12: Reaction time results: two-character words. Character 1 visual complexity (left panel) and character 2 visual complexity (right panel).

character frequency measures, the effect size for the visual complexity of the first character (64 ms) is larger than the effect size for the visual complexity of the second character (22 ms). Again, this indicates that lexical properties of the first character influence naming latencies to a greater extent than lexical properties of the second character.

For one-character words, properties of character components did not significantly influence naming latencies. The analysis for two-character words, by contrast, does show effects below the character level. These effects are limited to the components of the first character. As mentioned in Chapter 2, each character has a semantic radical. The two principal components that describe lexical properties of the semantic radical of the first character are `PC22` and `PC34`, which encode the frequency and visual complexity of the semantic radical of the first character.

The effect of `PC22` ($F = 22.636$, $p < 0.001$) is shown in the left panel of Figure 3.13. `PC22` has high positive loadings for `Character 1 SR Family Size` (i.e., the number of characters in which a semantic radical appears, loading: 0.918) and `Character 1 SR Frequency` (0.876). Therefore, a more frequent use of the semantic radical of the first character leads to longer naming latencies.

The effect of the frequency of the semantic radical is in the opposite direction of the word-level and character-level frequency effects described above. Furthermore, it is opposite to the effects of semantic radical family size in lexical decision experiments (Feldman & Siok, 1997, 1999b, 1999a). This pattern of results fits well with findings in English, where inhibitory effects of letter bigram frequency were observed in word naming (Hendrix, 2016). The opposite pattern of results for semantic radical frequency in lexical decision and word naming suggests that while high frequency semantic radicals help determine the lexical status of a character (i.e., real Chinese

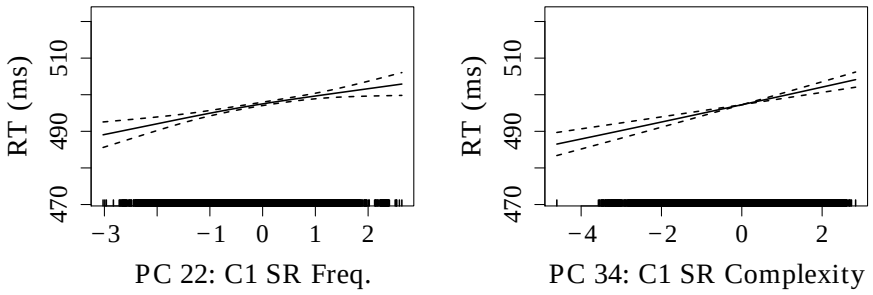

Figure 3.13: Reaction time results: two-character words. Character 1 SR frequency (left panel) and SR complexity (right panel).

character or not), they do not provide much information for the discrimination of a specific character.

The right panel of Figure 3.13 presents the effect for `PC34` ($F = 44.665$, $p < 0.001$). The only predictor with a high loading on `PC34` is `Character 1 SR Strokes`, with a loading of 0.923. All other predictors have loadings with absolute values smaller than 0.30 on `PC34`. The effect of the visual complexity of the semantic radical of the first character is similar to the effect of visual complexity at the character level, with longer naming latencies for more complex semantic radicals.

The effect sizes of the frequency and visual complexity of the semantic radical of the first character are limited, with an effect size of 14 ms for `PC22` and an effect size of 18 ms for `PC34`. Therefore, while measures of lexical properties of the semantic radical did reach significance for the first character, these effects are subtle. For the second character, we did not find significant effects of semantic radical measures at the Bonferroni-corrected α level.

In addition to the frequency effect of the semantic radical, we found an effect of the frequency of the high-level components in the first character, which are encoded in `PC16` (highest loadings: 0.951 for `Character 1 Max High-Level Component Frequency` and 0.944 for `Mean High-Level Component Frequency`; absolute values of all other loadings < 0.40). As can be seen in Figure 3.14, the effect of the frequency of the high-level components is similar to the effect of the frequency of the semantic radical, with longer naming latencies for word-initial characters with higher frequency high-level components ($F = 13.757$, $p < 0.001$). As was the case for the effect of the semantic radical frequency, the effect of the frequency of the high-level components is subtle, with an effect size of 12 ms.

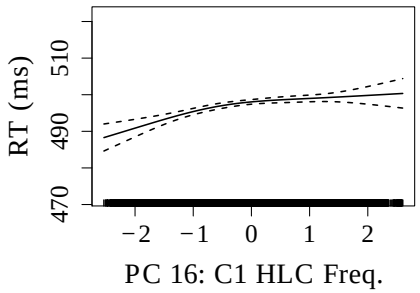

Figure 3.14: Reaction time results: two-character words. Character 1 high-level component frequency.

In comparison to the analysis for one-character words, the analysis for two-character words furthermore revealed two effects related to the pronunciation of a word. These effects are presented in Figure 3.15. The left panel of Figure 3.15 presents the effect of `PC12` ($F = 18.880$, $p < 0.001$). The lexical predictors that have high loadings on `PC12` are measures of the number of homographs of the first character: `Character 1 Homographs (Tokens)` (0.976), `Character 1 Homographs (Types)` (0.972) and `Character 1 Homographs Frequency` (0.956).

Despite the fact that we applied Yeo-Johnson power transformations to the input variables, the distribution of `PC12` is bimodal. As can be seen in the left panel of Figure 3.15, both parts of the distribution show an inhibitory effect. However, between the right edge of the left part of the distribution and the left part of the right edge of the distribution predicted values drop. This results in an overall effect of `PC12` that has a complicated non-linear form. Nonetheless, the overall trend of the effect is that a higher number of higher frequency homographs leads to longer naming latencies, with a difference of 33 ms between the predicted values for the lowest and highest values of `PC12`. Thus, the increased uncertainty about the pronunciation of the first character for characters with multiple pronunciations leads to additional processing costs.

In Chapter 2 we observed that first characters have fewer and less frequent homographs than second characters. We argued that for the first character the uncertainty about the character is relatively high, whereas for the second character the amount of uncertainty about the character is reduced through the information provided by the first character. The consistency between orthography and phonology thus is inversely proportional to the uncertainty about the character. On the basis of this observation, we would expect the processing costs for homography to be

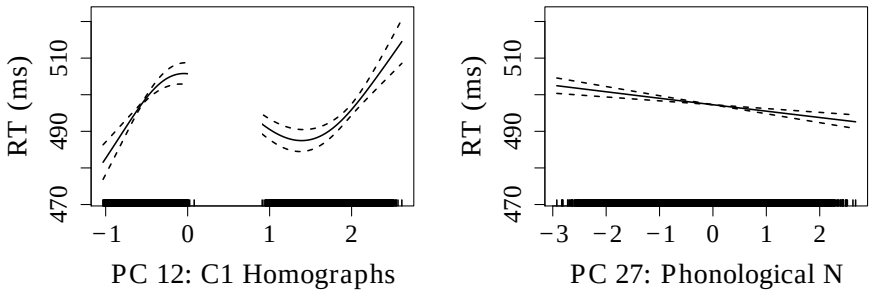

Figure 3.15: Reaction time results: two-character words. Character 1 homographs (left panel) and phonological neighbours (right panel).

higher for the first character than for the second character. This prediction is borne out: we observed an effect of homography for the first character, but not for the second character. For the second character, the additional processing costs due to homography thus are reduced by the information provided by the first character.

The right panel of Figure 3.15 shows the effect of the second principal component that encodes information about the pronunciation of a word, `PC27`, on which only `Phonological N` has a high loading (0.874; absolute values of all other loadings < 0.50). The effect of `PC27` ($F = 24.677$, $p < 0.001$) is facilitatory, with shorter naming latencies for words with more phonological neighbours. This effect is similar to the effects of phonological neighbourhood in the word naming task that were observed for English (see, e.g., Vitevich, 2002; Hendrix, 2016). However, the effect of phonological neighbourhood density is subtle, with a mere 10 ms difference between the predicted values for the highest and lowest values of `PC27`.

Thus far we reported effects of principal components that described lexical properties of the first character, the second character or the word as a whole. We now turn our attention to the effects of principal components that describe the relationship between the first and second character. First, we consider the effects of entropy. The entropy of the first and second character is described by `PC42` and `PC39`, respectively. The predictor with the highest loading on `PC42` is `Character 1 Entropy` (0.753), whereas the predictor with the highest loading on `PC39` is `Character 2 Entropy` (0.734). As a reminder, `Character 1 Entropy` is a measure of the uncertainty about the second character given a specific first character, whereas `Character 2 Entropy` is a measure of the uncertainty about the first character given a specific second character.

The effects of the entropy of the first (left panel) and second character (right panel) are presented in Figure 3.16. A higher entropy leads to shorter naming latencies, both for the first ($F = 134.006$, $p < 0.001$) and for the second character ($F = 41.597$, $p < 0.001$), with an effect size that is somewhat larger for the entropy of the first character (32 ms) than for the entropy of the second character (22 ms). The effect of entropy observed here is in the opposite of the entropy effects typically observed in English, which tend to show greater processing costs for high entropy items.

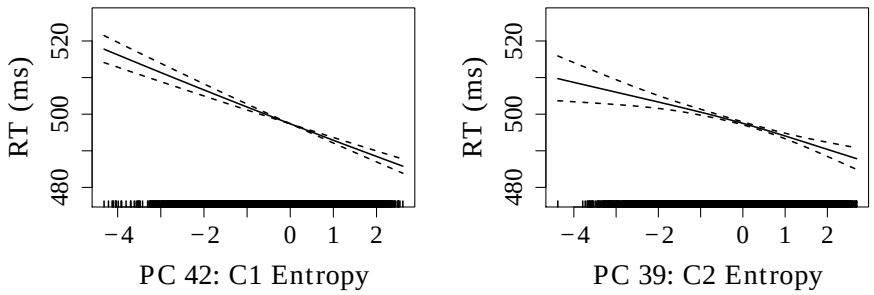

Figure 3.16: Reaction time results: two-character words. Character 1 entropy (left panel) and character 2 entropy (right panel).

One potential explanation of the facilitatory relative entropy effect is that the orthography-to-phonology mapping may be more consistent for characters that combine with many other characters as compared to characters that combine with few other characters. In this case, more uncertainty about the second character given the first character, and vice versa, would lead to less uncertainty about the phonological forms that need to be produced. An inspection of the loadings for `PC42` and `PC39` suggests that there may be some truth to this hypothesis. Unsurprisingly, the lexical predictors with the second highest loadings on `PC42` and `PC39` are `Character 1 Family Size` (0.259) and `Character 2 Family Size` (0.334). The lexical predictors with the third highest loadings on `PC42` and `PC39` are the number of friends for both characters.

The loading of `Character 1 Friends` on `PC42` is 0.255, whereas the loading of `Character 2 Friends` on `PC39` is 0.319. Therefore, the greater the uncertainty about the second character given the first character, the greater the number of words in which the same first character has the same pronunciation. Similarly, the greater the uncertainty about the first character given the second character, the greater the number of words in which the same second character has the same pronunciation.

A greater uncertainty about the identity of the other character is thus offset by a greater certainty about the pronunciation of the current character. The shorter naming latencies for words with high character one and character two entropies may therefore at least partially be a result of increased certainty about the pronunciation of these words.

A further entropy effect for two-character words is an effect of the trigram entropy of the first character, which is encoded in `PC50` (loading `Character 1 Trigram Entropy`: 0.682; no other loadings with absolute values greater than 0.20). The effect of `PC50` ($F = 16.397$, $p < 0.001$) is shown in Figure 3.17. The effect size of the first character trigram entropy effect is 19 ms. Naming latencies are shorter for first characters that appear in more character trigrams with more similar frequencies.

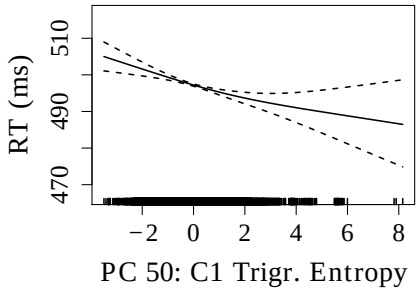

Figure 3.17: Reaction time results: two-character words. Character 1 trigram entropy.

One way to look at trigram entropy is as an equivalent of the entropy measure described above that operates both at and above the word level (correlation between `Character 1 Trigram Entropy` and `Character 1 Entropy`: $r = 0.687$). Alternatively, trigram entropy can be thought of as a more local alternative to contextual diversity (correlation between `Character 1 Trigram Entropy` and `Character 1 CD`: $r = 0.580$): more frequent characters tend to appear in a greater number of character trigrams. The facilitatory effect of trigram entropy, therefore, fits well with the facilitatory effects of both the entropy and contextual diversity of the first character reported above.

A third entropy effect is the effect of the relative entropy of the first character, which compares the similarity of the second character frequency distribution for the first character to the frequency distribution of all second characters in the CLD. The greater this similarity, the lower the relative entropy of the first character. The relative entropy of the first character is captured by `PC35`, which has a high loading

for `Character 1 RE` (0.971; absolute values of all other loading < 0.20) only. The left panel of Figure 3.18 presents the main effect of `PC35` ($F = 10.610$, $p < 0.001$). High values of relative entropy correspond to shorter naming latencies. The effect is most prominent for low predictor values and levels off for medium to high predictor values. The difference between that lowest and highest predictor values is 21 ms.

The overall facilitatory trend of the effect of the relative entropy of the first character is consistent with the effects of the entropy and trigram entropy of the first character. However, the GAM analysis revealed a significant interaction of `PC35` with `PC5` ($F = 6.457$, $p < 0.001$), which describes the frequency of the word as a whole.

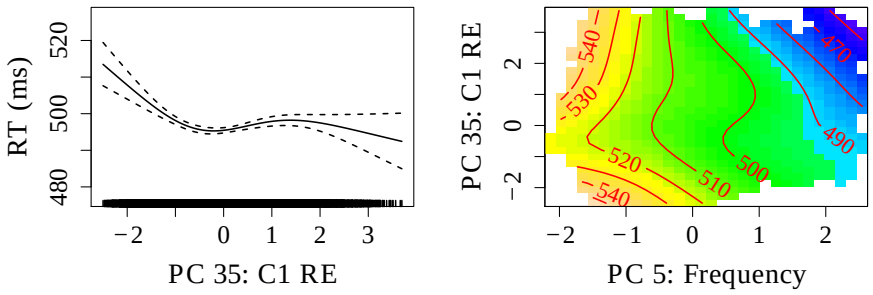

Figure 3.18: Reaction time results: two-character words. Character 1 RE (left panel) and the interaction of frequency with character 1 RE (right panel). Right panel shows the additive contour surface of the main effect of frequency, the main effect of character 1 RE and the interaction between frequency and character 1 RE.

The right panel of Figure 3.18 presents the additive contour surface of the main effect of `PC5`, the main effect of `PC35` and the interaction between `PC5` and `PC35`. For both high and low frequency words, naming latencies are relatively long when the relative entropy of the first character is low. Therefore, for both frequent and infrequent words, a prototypical frequency distribution of second characters results in additional processing time. A prototypical frequency distribution of second characters implies a relatively flat frequency distribution across character 1-character 2 combinations. In other words: a low value of relative entropy leads to more uncertainty about the identity of the second character. The relatively long naming latencies for first characters with a low relative entropy, as a result, reflects the increased difficulty of a choice problem.

For high frequency words, naming latencies are shorter for words with high values for PC35. However, for low frequency words, the effect of PC35 is U-shaped, with relatively long naming latencies for both low and high predictor values. Understanding this pattern of results requires a bit of thought about what it means for a character to have a high relative entropy. Above, we defined characters with low relative entropy as characters that have a relatively flat frequency distribution across character 1-character 2 combinations and thus reflect the overall frequency of those other characters across all two-character words reasonably well. By contrast, word-initial characters with a high relative entropy combine with relatively few second characters. The relative entropy furthermore increases when these few combinations have high frequencies. A character with a high relative entropy, therefore, is best described as a character that forms high-frequency two-character words with only a few second characters.

For high frequency words with a high relative entropy, the current word is (one of) the high frequency combination(s) of the first character with a second character that resulted in a high value of relative entropy. Given the sparsity or even absence of other high frequency words that share the same first character, there is a strong expectation that the second character will appear when the first character is read, which turns out to be correct and facilitates lexical processing.

For low frequency words with a high relative entropy the situation reverses. The combination of the current first and second character is uncommon. In addition, there are one or more high frequency words with the same first character. When reading the first character, therefore, there is a strong expectation that the first character will be combined with a second character to form one of these high frequency words. This prediction turns out to be incorrect, which leads to additional processing costs.

The fourth entropy effect is the effect of PC36 ($F = 18.298$, $p < 0.001$). The lexical predictor with the highest loading on PC36 is Entropy Character Frequencies (0.958; absolute values of all other loadings < 0.20). The entropy over the character frequencies is a measure of the similarity of the frequency of the left and right character. The main effect of PC36 is presented in the left panel of Figure 3.19, which shows that naming latencies are 21 ms longer when the characters in a two-character word have a similar frequency.

However, in addition to a main effect of PC36, we also observed an interaction between PC36 and PC1 ($F = 13.377$, $p < 0.001$). PC1, as a reminder, encodes the frequency of the first character. The interaction is presented as an additive contour

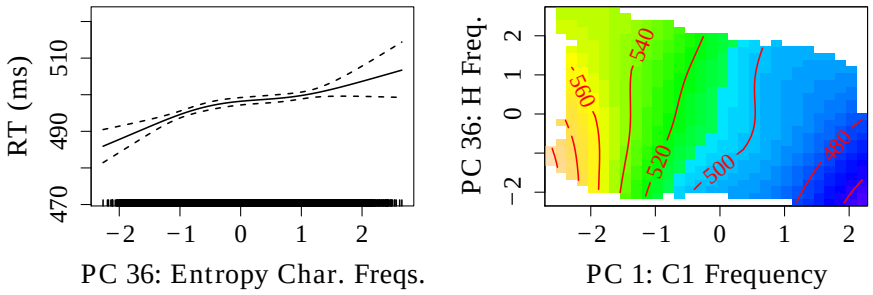

Figure 3.19: Reaction time results: two-character words. Entropy character frequencies (left panel) and the interaction of character 1 frequency with entropy character frequencies (right panel). The right panel shows the additive contour surface of the main effect of character 1 frequency, the main effect of entropy character frequencies and the interaction between character 1 frequency and entropy character frequencies.

plot of the main effects of both predictors and their interaction in the right panel of Figure 3.19.

For words with frequent first characters, the effect of the entropy over the character frequencies corresponds to the main effect of `PC36`, with longer naming latencies for words that consist of characters with similar frequencies. For these words the frequency of both the first and the second character is high. Given that frequency of a character correlates strongly with its family size (`Character 1 Family Size` has a loading of 0.862 on `PC1`), this implies that both the first and the second character combine with many other characters to form two-character words. That is, the uncertainty about the identity of the word given either the first or second character is high for these words.

By contrast, for words with infrequent first characters a high entropy over the character frequencies implies that not only the first character, but also the second character has a low frequency. As was the case for the first character, the frequency of the second character correlates strongly with its family size (`Character 2 Family Size` has a loading of 0.830 on `PC2`, which describes the frequency of the second character). Both the first and the second character thus combine with a limited number of other characters to form two-character words. As a result, the uncertainty about the identity of the word as a whole is limited. Hence, although naming latencies are still relatively long due to a low first character frequency, they are decreased as compared to words with a low frequency first character and a higher frequency second character.

Finally, we observed an effect of the categorical variable `Character 2 Type`. An investigation of the pairwise differences revealed that only the difference between pictophonetic and pictographic characters reached significance at the Bonferroni-corrected α level (t = -5.414, $p < 0.001$), with longer naming latencies for words in which the second character was a pictophonetic character as compared to words in which the second character was a pictographic character.

Pictophonetic characters are the only character type for which a phonetic radical is present. However, the presence of a phonetic radical does not help pronounce words faster. Instead, naming latencies are slower when the second character contains a phonetic radical. The inhibitory effect of the presence of a phonetic radical may be a result of the fact that phonetic radicals provide less-than-reliable information about the pronunciation of a character. More often than not (see the introduction of the categorical variables in Chapter 2), the pronunciation of the phonetic radical is different from the pronunciation of the character it occurs in. Not only do phonetic radicals therefore add additional information that needs to be processed to a character, this information may also provide misleading information about the pronunciation of a word.

3.4.2 Pronunciation durations

3.4.2.1 GBM

The relative influences of the 153 predictors in a GBM fitted to the pronunciation durations are shown in Table 3.2. As was the case for the naming latencies, the control variables have substantial contributions to the pronunciation durations, with high relative influences for `Initial Phoneme` (11.970%), `Final Phoneme` (2.695%), `Session` (11.126%) and Trial (1.333%).

Figure 3.20 presents the summed relative influence for the clusters and groups of predictors. The total summed relative influence of the control variables is 27.124%. Figure 3.20 furthermore indicates that over half of the relative influence of predictors in the GBM is accounted for by measures of phonological properties of the word and its characters. In total, the predictors in the 6 clusters of phonological measures have a summed relative influence of no less than 54.097%.

A large part of the explanatory power for the measures in the group of phonological predictors (Group 3) comes from measures describing the phonological frequency of both characters. The strongest predictor in the GBM is `Character 2 Mean Phoneme Frequency`, with a relative influence of 32.193%. Five further measures of

Table 3.2: Relative variable influences in an XGBoost model fitted to the pronunciation durations. Abbreviations: BE = Backward Enemies, C1 = Character 1, C2 = Character 2, Diph. = Diphone, Freq. = Frequency, HLC = High-Level Components, LLC = Low-Level Components, Phono = Phonological, Phon. = Phoneme, SUBTL = SUBTLEX-CH, Typ. = Types, Tok. = Tokens.

rank	predictor	infl.	rank	predictor	infl.
1	C2 Mean Phon. Freq.	32.193	29	Mean Diph. Freq.	0.145
2	C2 Phono N	13.652	30	Frequency (Gigaword)	0.140
3	Initial Phoneme	11.970	31	C1 Mean Phon. Freq.	0.131
4	Session	11.126	32	C1 PLD	0.130
5	C2 Tone	4.697	33	C2 Homograph (Tok.)	0.122
6	C2 Mean LLC Freq.	3.015	34	C1 Phono N	0.119
7	Final Phoneme	2.695	35	C2 Homophones (Typ.)	0.118
8	C2 Picture Size	2.505	36	C1 Min Diph. Freq.	0.111
9	C1 Tone	2.322	37	C1 Mean Diph. Freq.	0.109
10	Trial	1.333	38	Mean Phon. Freq.	0.101
11	C2 Min Phon. Freq.	1.252	39	Min Phon. Freq.	0.100
12	C2 CD (Gigaword)	1.066	40	Max Diph. Freq.	0.087
13	C2 Initial Phon. Freq.	1.028	41	C1 Freq. (SUBTL)	0.087
14	C2 Min HLC Freq.	0.902	42	Min Diph. Freq.	0.082
15	C2 Max Phon. Freq.	0.761	43	C2 RE	0.081
16	C1 Max Diph. Freq.	0.747	44	C2 Phonemes	0.081
17	Phonemes	0.655	45	C2 Friends	0.079
18	C2 Strokes	0.588	46	C2 Trigram Entropy	0.079
19	C2 Initial Diph. Freq.	0.511	47	Frequency (SCCoW)	0.079
20	C2 Mean Diph. Freq.	0.490	48	t-Score	0.070
21	C2 Min Diph. Freq.	0.390	49	PLD	0.068
22	Transitional Diph. Freq.	0.286	50	Entropy Char. Freqs.	0.062
23	C1 Phonemes	0.275	51	C1 Trigram Entropy	0.061
24	C2 Max Diph. Freq.	0.209	52	C1 PR Frequency	0.059
25	Phono N	0.195	53	C2 Mean HLC Freq.	0.056
26	C2 Homophones Freq.	0.177	54	C2 PR BE (Tok.)	0.055
27	CD (SUBTL)	0.167	55	C1 Picture Size	0.053
28	Frequency (SUBTL)	0.165	56	C1 Entropy	0.051

Table 3.2 (continued)

rank	predictor	infl.	rank	predictor	infl
57	C1 Pixels	0.051	90	PMI	0.029
58	C1 RE	0.050	91	C1 PR Strokes	0.029
59	Max Phon. Freq.	0.050	92	C1 Min LLC Freq.	0.029
60	C1 Homograph (Tok.)	0.049	93	C2 Freq. (SUBTL)	0.027
61	C1 Pixels OLD	0.049	94	C1 Freq. (SCCoW)	0.027
62	C2 Min LLC Freq.	0.048	95	C1 Traditional Freq.	0.027
63	C1 Mean LLC Freq.	0.047	96	C1 Initial Phon. Freq.	0.025
64	C2 Family Size	0.047	97	C2 SR Strokes	0.025
65	C1 Max Phon. Freq.	0.046	98	C2 CD (SUBTL)	0.024
66	C1 CD (SUBTL)	0.045	99	C1 PR Enemies Freq.	0.024
67	Position-specific PMI	0.043	100	Traditional Freq.	0.024
68	CD (Gigaword)	0.042	101	C1 LLC OLD	0.024
69	C2 Pixels OLD	0.042	102	C1 Family Freq.	0.022
70	C1 CD (SCCoW)	0.041	103	C1 Homophones Freq.	0.022
71	C2 PR Frequency	0.040	104	C2 Homographs Freq.	0.021
72	C1 Min Phon. Freq.	0.040	105	C1 PR Enemies (Tok.)	0.021
73	C1 PR Friends Freq.	0.037	106	C1 PR BE Freq.	0.020
74	C1 Friends Freq.	0.036	107	C2 PR Enemies (Tok.)	0.020
75	C1 Family Size	0.036	108	C1 Homographs Freq.	0.018
76	C2 LLC OLD	0.036	109	C1 PR Friends	0.018
77	C1 Homophones (Typ.)	0.033	110	C1 Freq. (Gigaword)	0.018
78	C1 Min HLC Freq.	0.033	111	C1 LLC	0.017
79	C1 Homophones (Tok.)	0.032	112	C2 CD (SCCoW)	0.017
80	C2 Max HLC Freq.	0.032	113	C2 Traditional Freq.	0.017
81	C2 Family Freq.	0.032	114	C2 PR Friends	0.017
82	C1 SR Frequency	0.032	115	C1 Initial Diph. Freq.	0.017
83	C2 PR BE Freq.	0.032	116	C1 Friends	0.016
84	CD (SCCoW)	0.032	117	C2 Entropy	0.015
85	C1 Mean HLC Freq.	0.032	118	C1 Strokes	0.015
86	C2 Homophones (Tok.)	0.030	119	C2 SR Frequency	0.013
87	C2 Pixels	0.030	120	Strokes	0.013
88	C2 Friends Freq.	0.030	121	C1 PR Family Size	0.012
89	C2 Freq. (SCCoW)	0.029	122	C1 Max HLC Freq.	0.012

Table 3.2 (continued)

rank	predictor	infl.	rank	predictor	infl.
123	C2 SR Family Size	0.011	139	C2 HLC	0.004
124	C2 PR Strokes	0.011	140	C2 PR Enemies (Typ.)	0.003
125	C2 PR Family Size	0.011	141	C1 Structure	0.003
126	C2 PLD	0.011	142	C2 LLC N	0.003
127	C1 PR BE (Typ.)	0.010	143	C2 Homographs (Typ.)	0.002
128	C2 PR BE (Typ.)	0.010	144	C2 Max LLC Freq.	0.002
129	C1 SR Strokes	0.009	145	C1 PR Regularity	0.001
130	C2 PR Enemies Freq.	0.009	146	C2 PR Regularity	0.001
131	C1 PR BE (Tok.)	0.008	147	C1 LLC N	0.001
132	C1 CD (Gigaword)	0.008	148	C1 Type	0.001
133	C1 PR Enemies (Typ.)	0.006	149	Length	0.000
134	C2 Freq. (Gigaword)	0.006	150	C1 HLC	0.000
135	C1 SR Family Size	0.006	151	C1 Max LLC Freq.	0.000
136	C2 PR Friends Freq.	0.006	152	C1 Homographs (Typ.)	0.000
137	C2 Structure	0.006	153	C2 Type	0.000
138	C2 LLC	0.004			

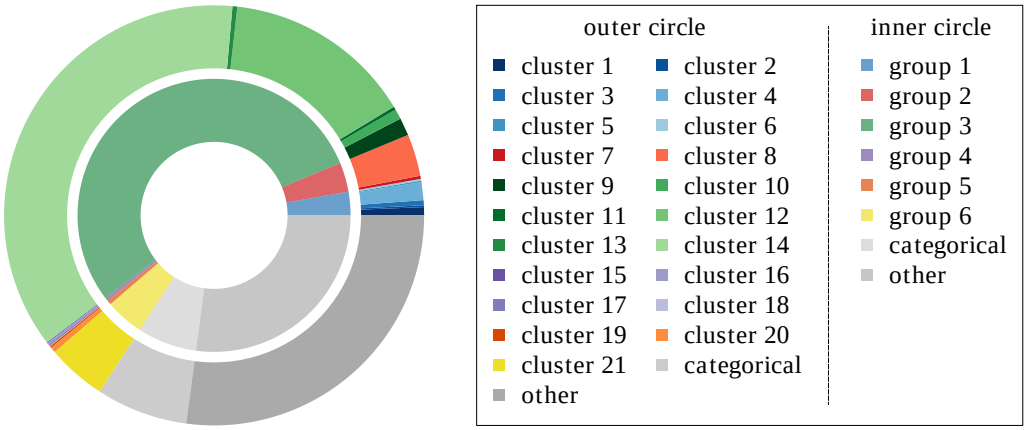

Figure 3.20: Relative variable influence per cluster (outer circle) and per group of clusters (inner circle) in an XGBoost model fitted to the pronunciation durations.

the frequency of the second character are among the 20 predictors with the highest relative influence. `Character 2 Min Phoneme Frequency` (1.252%), `Character 2 Initial Phoneme Frequency` (1.028%), `Character 2 Max Phoneme Frequency` (0.761%), `Character 2 Initial Diphone Frequency` (0.511%), and `Character 2 Mean Diphone Frequency` (0.490%) all contribute to the explanatory power of the GBM model, although to a much lesser extent than `Character 2 Mean Phoneme Frequency`.

In total, the two clusters that contain measures of the phonological frequency of the second character (Cluster 14 and Cluster 10) have a summed relative influence of 37.534%. By comparison, the measures in the two clusters containing measures that describe the phonological frequency of the first character (Cluster 13 and Cluster 9) have a summed relative influence of 1.733%. The only measure of the phonological frequency of the first character among the 20 predictors with the highest relative influence is `Character 1 Max Diphone Frequency`, which has a relative influence of 0.747%. The phonological frequency of the second character, therefore, has a much greater influence on pronunciation durations than does the phonological frequency of the first character.

In addition to phonological frequencies, phonological neighbourhood characteristics influence pronunciation durations as well. As was the case for the phonological frequency measures, phonological neighbourhood density measures for the second character (summed relative influence Cluster 12: 14.581%) have greater explanatory power as compared to phonological neighbourhood density measures for the first character (summed relative influence Cluster 11: 0.249%). The phonological neighbourhood density measure with the greatest relative influence is `Character 2 Phonological N`. With a relative influence of 13.652%, this predictor is the second best predictor for pronunciation durations, after `Character 2 Mean Phoneme Frequency`.

A further measure that was grouped with the phonological density measures for the second character in our clustering analysis on the SOM is `Phonemes` (i.e., the number of phonemes in a word). With a relative influence of 0.655%, `Phonemes` was among the 20 predictors with the highest relative influence as well. Although relatively mild as compared to the explanatory power of the phonological frequency and phonological neighbourhood density measures, phonological complexity thus has an influence on pronunciation durations as well.

The GBM fitted to the pronunciation durations furthermore attributed some importance to visual complexity measures (summed relative influence for the measures in the cluster in Group 2: 3.475%). Again, the summed relative influence of the second character measures (Cluster 8, 3.236%) is greater than the summed relative influence of the first character measures (Cluster 7, 0.239%). As can be seen in Table 3.2, two measures of the visual complexity of the second character are among the 20 predictors with the greatest relative influence: `Character 2 Picture Size` (2.505%) and `Character 2 Strokes` (0.588%).

In the GBM for the naming latencies, the group of clusters with the strongest predictive power was Group 1, which describes the frequency of a word and its characters. As can be seen in Figure 3.20, the role of frequency-related measures in the GBM fitted to the pronunciation durations is much smaller. In total, the 6 clusters that consist of frequency measures have a summed relative influence of a mere 2.784%. The only frequency-related measure in the top 20 of predictors with the highest relative influence is `Character 2 CD (Gigaword)` (relative influence: 1.066%).

The final two numerical predictors that contributed considerably to the explanatory power of the GBM fitted to the pronunciation durations are measures of the frequency of the high-level and low-level components of the second character, which are part of Cluster 21 (summed relative influence: 4.565%). These two measures are `Character 2 Mean Low-Level Component Frequency` (relative influence: 3.015%) and `Character 2 Min High-Level Component Frequency` (0.902%). Pronunciation durations, therefore, are co-determined to a greater extent by component frequencies as compared to character and word frequencies. However, compared to the effects of phonological lexical properties, the contribution of orthographic frequency is limited at all grain sizes.

The last two groups of numerical predictors, Group 4 and Group 5 contributed little to the explanatory power of the GBM. The summed relative influence of the homograph measures in Group 4 was 0.377%, whereas the summed relative influence of the homophone measures in Group 5 was 0.547%. The impact of the consistency between orthography and phonology on pronunciation durations is therefore limited.

Finally, none of the categorical predictors in Group 7 was among the 20 predictors with the highest relative influence for the GBM fitted to the naming latencies. By contrast, two categorical variables were in the top 20 predictors with the highest relative influence for the pronunciation durations: `Character 1 Tone` and `Char-`

acter 2 Tone. Consistent with our observations above, the relative influence of Character 2 Tone (4.697%) was greater than the relative influence of Character 1 Tone (2.322%). Together, Character 1 Tone and Character 2 Tone account for most of the summed relative influence for the categorical variables, which is 7.031%

In conclusion, the GBM analysis indicates that the lexical predictors with the strongest influence on pronunciation durations are measures describing phonological properties of the word and its characters. Additionally, we observed smaller contributions of visual complexity and orthographic frequency measures. Overall, lexical properties of the second character had substantially greater relative influences than lexical properties of the first character.

3.4.2.2 GAM

3.4.2.2.1 One-character words

A GAM fitted to the pronunciation durations for one-character words showed significant effects of the control variables Initial Phoneme ($F = 99.993$, $p < 0.001$) and Final Phoneme ($F = 87.324$, $p < 0.001$). Furthermore, we observed effects of Session ($F = 82.231$, $p < 0.001$) and Trial ($F = 16.359$, $p < 0.001$). Figure 3.21 presents the effects of Session (left panel) and Trial (right panel).

Pronunciation durations are shorter for high predictor values for both predictors. The effect of Session is linear, with predicted pronunciation durations being 20 ms longer in the first experimental session as compared to the last experimental session. The effect of Trial is non-linear. Pronunciation durations decrease over the course of the first 300 trials in an experimental session. After 300 trials, the effect of Trial levels off and pronunciation durations no longer decrease. The difference between the longest and shortest predicted pronunciation durations is 19 ms.

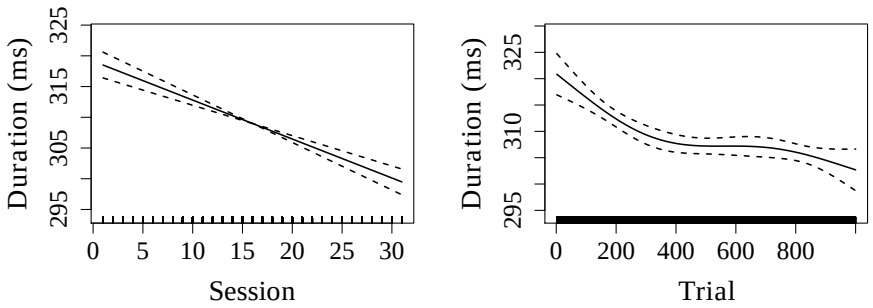

Figure 3.21: Duration results: one-character words. Experimental predictors.

Despite the limited relative influence of frequency measures in the GBM analysis of the pronunciation durations, we furthermore found a moderate effect of PC1 ($F = 18.090$, $p < 0.001$), which encodes the frequency of the character and the word. The effect of PC1 is presented in Figure 3.22. Pronunciation durations are 16 ms shorter for high frequency word and/or characters as compared to low frequency words and/or characters. The effect is linear for low values of PC1 and levels off for high values of PC1 (c.f., Jurafsky et al. (2001) for a discussion of similar reduction effects in English). Most likely, the effect of PC1 is a learning effect. The shorter pronunciation durations for high frequency words then reflect the additional experience in pronouncing these words.

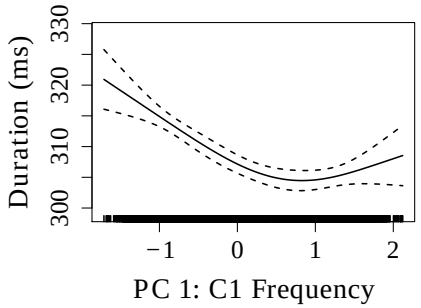

Figure 3.22: Duration results: one-character words. Character frequency.

In addition to the effect of frequency at the character and the word level, we observed two effects of frequency below the word level. Consistent with the results of the GBM analysis, both of these effects are phonological frequency effects. First, we observed an effect of PC8 ($F = 22.282$, $p < 0.001$). The lexical predictors with the highest loadings on PC8 are Character 1 Max Phoneme Frequency and Character 1 Mean Phoneme Frequency, with loadings of 0.932 and 0.862. Second, we found an effect of the diphone counterpart of PC8, PC6 ($F = 21.867$, $p < 0.001$). The lexical predictors with the highest loadings on PC6 are Character 1 Max Diphone Frequency (0.978) and Character 1 Mean Diphone Frequency (0.919).

The effects of PC8 (left panel) and PC6 (right panel) are shown in Figure 3.23. For both principal components, the confidence intervals for extreme predictor values are wide. However, for non-extreme predictor values, the effects of both phoneme frequency and diphone frequency are inhibitory, with higher frequency phonemes and diphones leading to longer pronunciation durations. If we interpret longer pronunciation durations as an indication of additional processing costs, the inhibitory

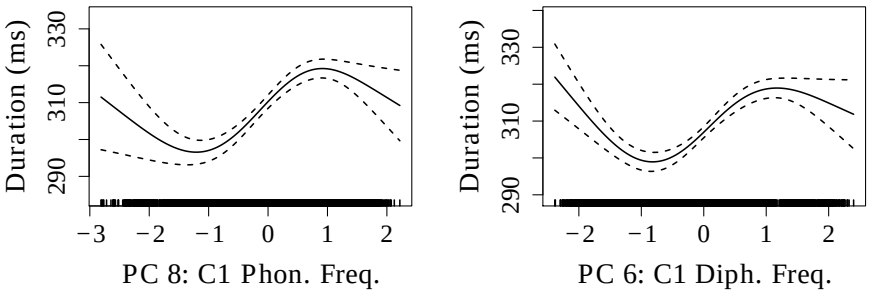

Figure 3.23: Duration results: one-character words. Character phoneme frequency (left panel) and diphone frequency (right panel).

effects of these sub-character level phonological frequency measures are in line with the inhibitory effects of sub-character level orthographic frequency measures in the GAM analysis of the naming latencies. The effects of `PC8` and `PC6` have similar effect sizes. The difference between the highest and lowest predicted values in the area of Figure 3.23 where the confidence intervals are narrow is 23 ms for `PC8` and 20 ms for `PC6`.

The frequency effects described thus far are complemented by an effect of phonological complexity, as encoded in `PC19` (loading `Character 1 Phonemes`: 0.739, absolute values of all other loadings < 0.30). The effect of `PC19` ($F = 19.763$, $p < 0.001$) is presented in Figure 3.24. Unsurprisingly, pronunciation durations are longer for words with a greater number of phonemes. The effect is linear and has an effect size of 29 ms. Note, however, that the p-value for the effect of `PC19` drops below the Bonferroni-corrected α level (0.000042) when the model is refitted to the subset of the data for which model residuals are within 2.5 standard deviations from the residual mean ($F = 15.584$, $p = 0.000080$).

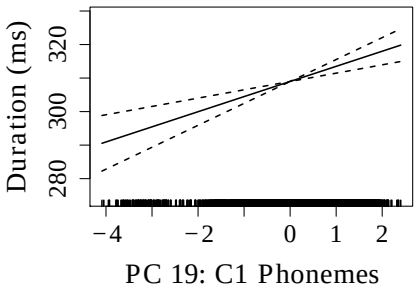

Figure 3.24: Duration results: one-character words. Character phonemes.

Finally, we observed an effect of the categorical variable `Character 1 Tone`. Consistent with previous studies on the duration of tones (Yu, 2010; Xu, 1997; Ho, 1976; Howie, 1974) predicted pronunciation durations are longest for Tone 3 (+82 ms), followed by Tone 2 (+33 ms), Tone 5 (+20 ms), Tone 1 (reference level) and Tone 4 (−20 ms). All pairwise comparisons between the 5 tones were significant at the Bonferroni-corrected α level, with the exception of the comparisons between Tone 5 and Tone 1 and Tone 5 and Tone 2.

3.4.2.2.2 Two-character words

As was the case for the GAM analysis for one-character words, the GAM analysis for two-character words showed significant effects of `Initial Phoneme` ($F = 324.999$, $p < 0.001$) and `Final Phoneme` ($F = 196.077$, $p < 0.001$). In addition, we observed effects of `Session` ($F = 2869.076$, $p < 0.001$) and `Trial` ($F = 227.723$, $p < 0.001$). The main effects of `Session` (left panel) and `Trial` (right panel) are presented in Figure 3.25.

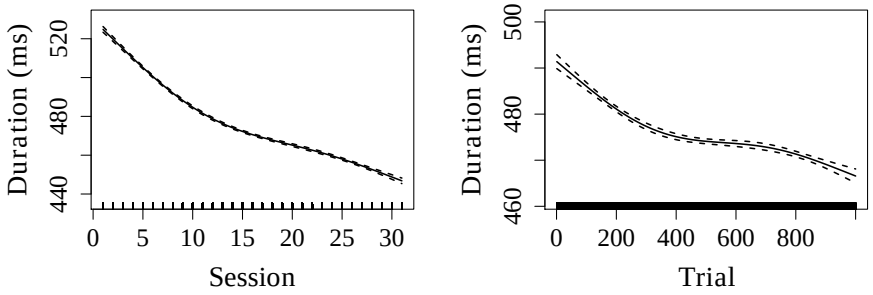

Figure 3.25: Duration results: two-character words. Session (left panel) and Trial (right panel).

The main effects of `Session` and `Trial` are qualitatively similar to the main effects of these predictors for one-character words, with a near-linear facilitatory effect for `Session` and a facilitatory effect for `Trial` that levels off for high predictor values. However, at 78 ms, the effect size of `Session`, is much larger for two-character words than for one-character words (20 ms). The effect size for the effect of `Trial` is larger for two-character words (25 ms) than for one-character words (19 ms) as well. However, the difference in effect size is much less pronounced as compared to the difference in the effect size for one-character words and two-character words for `Session`.

In addition to the main effects of `Session` and `Trial`, we observed an interaction between both predictors ($F = 7.501$, $p < 0.001$). Figure 3.26 presents the interaction between `Session` and `Trial`. The effect of `Trial` is most prominent for the early experimental sessions: the difference between the highest and the lowest predicted values for the first experimental session is 34 ms, whereas the difference between the highest and the lowest predicted values for the last experimental session is 13 ms.

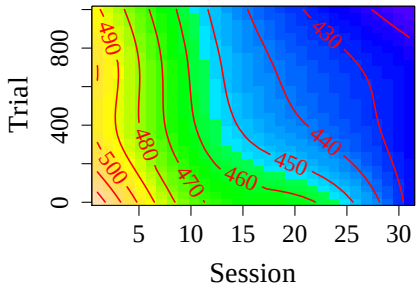

Figure 3.26: Duration results: two-character words. Interaction between Session and Trial. Additive contour surface for the main effect of Session, the main effect of Trial and the interaction between Session and Trial.

For one-character words, we found an effect of `PC1`. Due to the high correlation between character frequency and word frequency for one-character words, `PC1` for one-character words encodes both character and word frequency. For two-character words, the frequencies at the word and character level are much less correlated. As a result, character frequency and word frequency allocated to separate principal components for two-character words. For two-character words we are therefore able to establish whether or not the frequency effect on pronunciation durations is primarily an effect of character frequency or an effect of word frequency.

The GAM for two-character words revealed no effects of character frequency. By contrast, we did observe effects of the word frequency measures `PC5` (word frequency in the SCCoW and in the Gigaword corpus) and `PC25` (word frequency in SUBTLEX-CH). The effects of `PC5` ($F = 65.601$, $p < 0.001$) and `PC25` ($F = 53.320$, $p < 0.001$) are presented in Figure 3.27 and demonstrate that the effect of frequency on pronunciation durations is primarily an effect of word frequency, rather than an effect of character frequency. This reflects the fact that character-level information needs to be integrated at the word level for articulatory planning.

The left panel of Figure 3.27 shows the effect of PC5, the frequency of the word in the SCCoW and in the Gigaword corpus. As was the case for one-character words, pronunciation durations are shorter for two-character words with a higher frequency. Whereas the effect for one-character words had a non-negligible non-linear component, the effect of PC5 is near-linear. The right panel of Figure 3.27 shows the effect of PC25, which encodes the frequency of the word in SUBTLEX-CH. Again, pronunciation durations are shorter for high predictor values. Consistent with the results of the GAM analysis for one-character words, the effect sizes for the principal components encoding word frequency in the SCCoW and Gigaword corpus (16 ms) on the one hand and in SUBTLEX-CH (17 ms) on the other hand are highly similar. Hence, the SUBTLEX-CH frequencies have similar explanatory power as compared to the frequencies from the SCCoW and the Gigaword corpus.

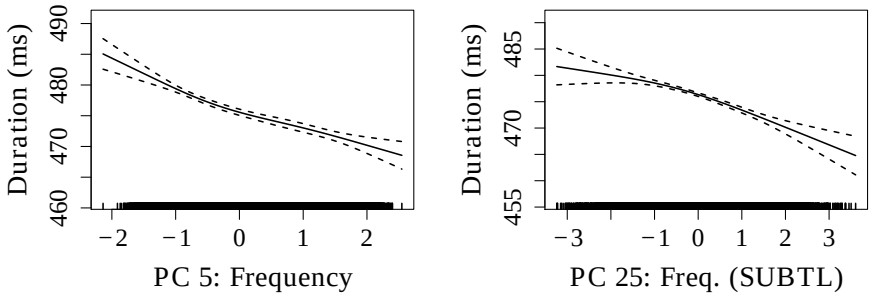

Figure 3.27: Duration results: two-character words. Word frequency (left panel) and word frequency in the SUBTLEX-CH corpus (right panel).

The GAM analysis for one-character words did not reveal any effects of visual complexity. Consistent with the GBM on the pronunciation durations we did observe an effect of visual complexity for the second character in the GAM analysis for two-character words. In particular, we found an effect of PC45 ($F = 11.919$, $p < 0.001$). The lexical predictor with the highest loading in PC45 is Character 2 Picture Size (0.775; absolute values of loadings for all other predictors < 0.20), which is a measure of the amount of information in a character.

The effect of PC45 is shown in Figure 3.28. As a result of the wide confidence intervals for all but medium predictor values considerable uncertainty remains about the qualitative nature of the effect of PC45. For medium predictor values Figure 3.28 pronunciation durations show somewhat of a decrease as pictures contain more information. Processing may therefore be optimized for the typical amount of

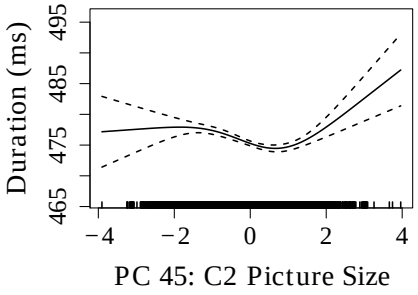

Figure 3.28: Duration results: two-character words. Character 2 picture size.

information in the visual input. However, this effect is extremely subtle, with an effect size smaller than 5 ms.

In addition to the frequency effect at the word level, the GAM analysis for two-character words revealed a series of frequency effects below the character level. For one-character words, we found an effect of phoneme frequency. For two-character words, we similarly observed a phoneme frequency effect. However, this effect was limited to the phoneme frequency of the second character, as encoded in `PC6` (highest loadings for `Character 2 Max Phoneme Frequency` (0.954) and `Character 2 Mean Phoneme Frequency` (0.902)).

The effect of `PC6` ($F = 135.753$, $p < 0.001$; effect size: 20 ms) is presented in Figure 3.29. The effect is inverse U-shaped. Pronunciation durations are longest for second character with medium-to-high phoneme frequencies and shortest for second characters with phoneme frequencies at the extremes of the predictor range. For one-character words, we observed a similar effect. However, the confidence intervals for one-character words were wide, such that definite conclusions could only be drawn about the increase of pronunciation durations in the middle part of the predictor range.

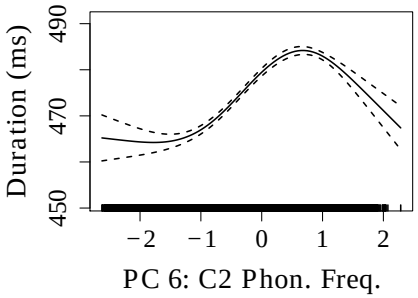

Figure 3.29: Duration results: two-character words. Character 2 phoneme frequency.

The increase in pronunciation durations in the lower part of the second character phoneme frequency range is in line with the effects of sub-character level frequencies reported for the naming latencies. From a learning perspective (see e.g., Baayen et al., 2011), this inhibitory effect makes sense. Learning theory predicts that the association between a word and its phonemes is inversely proportional to the number of words in which these phonemes occur. For characters with low frequency phonemes the association between the character and its phonemes thus is strong. This results in easier access to the target phonemes and shorter pronunciation durations.

The decrease in pronunciation durations for high predictor values is inconsistent with such an interpretation – and suggests that there may be a trade-off between learning and articulatory fluency. On the one hand, the associations between high frequency phonemes and the characters they occur in are weak. One the other hand, speakers have more experience pronouncing high frequency phonemes as compared to low frequency phonemes. The current effect of phoneme frequency suggests that for words with high frequency phonemes, the benefit of increased articulatory fluency is greater than the cost of decreased association strengths. Pronunciation durations thus are relatively short when the pronunciation of the second character consists of highly frequent phonemes.

In addition to the effect of mean (and maximum) phoneme frequency, the GAM revealed effects of the minimum phoneme frequency of both the first and the second character. The minimum phoneme frequency of the first character is encoded in `PC14`. The lexical predictors with the highest loadings on `PC14` are `Character 1 Initial Phoneme Frequency` (0.930) and `Character 1 Min Phoneme Frequency` (0.870). `PC17` describes the minimum phoneme frequency of the second character. The predictors with the highest loadings on `PC17` are `Character 2 Initial Phoneme Frequency` (0.890) and `Character 2 Min Phoneme Frequency` (0.811). Both the minimum phoneme frequency and the initial phoneme frequency of the first and second character have high loadings on `PC14` and `PC17`, respectively. This is due to the fact that the initial phoneme of a character is often the least frequent phoneme, which is the case for 81.30% of the first characters and 81.61% of the second characters in the all two-character words in the CLD.

Figure 3.30 presents the effects of the minimum phoneme frequency for the first (left panel) and the second character (right panel). The effect of the minimum phoneme frequency for the first character (`PC14`; $F = 9.661$, $p < 0.001$) is U-shaped with longer pronunciation durations for first characters with non-typical minimum

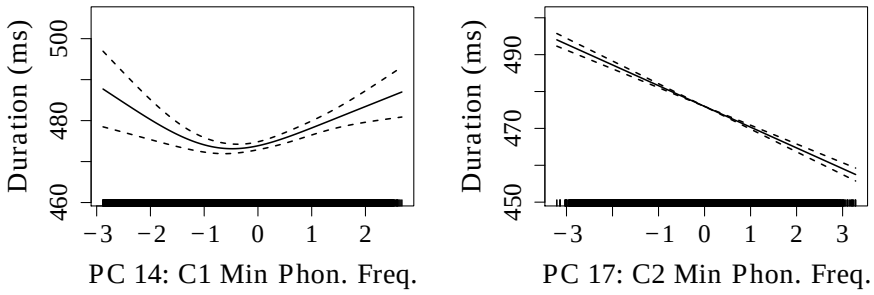

Figure 3.30: Duration results: two-character words. Minimum phoneme frequency for character 1 (left panel) and character 2 (right panel).

phoneme frequencies. However, at 15 ms, the effect size of the minimum phoneme frequency effect for character 1 is limited. Furthermore, the confidence intervals are wide for large parts of the predictor range. Therefore, the evidence for a U-shaped effect of the minimum phoneme frequency of the first character is weak.

The GAM analysis provided much stronger evidence for an effect of the minimum phoneme frequency of the second character (`PC17`; $F = 443.405$, $p < 0.001$). Pronunciation durations are shorter for high predictor values. The effect is linear in nature and has an effect size of 37 ms. For high values of minimum phoneme frequency the association between the phoneme with the minimum frequency and the character is relatively weak (expectation: longer pronunciation durations), whereas the acoustic fluency is relatively high (expectation: shorter pronunciation durations). The shorter pronunciation durations for high values of `PC17` thus suggest that the effect of `PC17` is primarily driven by the increased acoustic fluency for second characters that do not contain low frequency phonemes.

In addition to effects of phoneme frequency, the GAM analysis revealed effects of the diphone frequency of both the first and the second character. The diphone frequency of the first character is encoded in `PC7` (lexical predictors with the highest loadings: `Character 1 Max Diphone Frequency` (0.981) and `Character 1 Mean Diphone Frequency` (0.923)), whereas `PC8` describes the diphone frequency for the second character (predictors with the highest loadings: `Character 2 Max Diphone Frequency` (0.979) and `Character 1 Mean Diphone Frequency` (0.921)). The effects for `PC7` ($F = 184.296$, $p < 0.001$) and `PC8` ($F = 26.589$, $p < 0.001$) are presented in Figure 3.31.

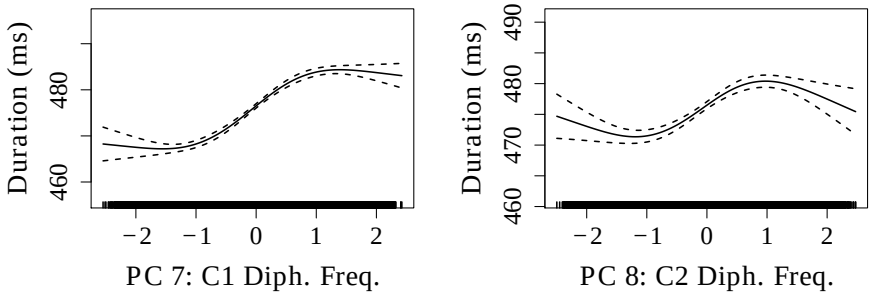

Figure 3.31: Duration results: two-character words. Diphone frequency for character 1 (left panel) and character 2 (right panel).

As can be seen in Figure 3.31, the effects of `PC7` (left panel) and `PC8` (right panel) are qualitatively similar to the effect of diphone frequency for one-character words. While the wide confidence intervals for extreme predictor values lead to uncertainty near the edges of the predictor range, we see a more reliable effect for non-extreme predictor values: pronunciation durations are longer for characters that consist of more frequent diphones. As before, this effect may reflect reduced associations with the character for high frequency phonological units as compared to low frequency phonological units. Nonetheless, at 17 and 9 ms, the diphone frequency effects for the first and second character are subtle.

A final frequency effect below the character level is an effect of the minimum diphone frequency of the word as a whole ($F = 46.367$, $p < 0.001$). The lexical predictors with the highest loadings on the corresponding principal component, `PC20`, are `Transitional Diphone Frequency` (0.872) and `Character 2 Initial Phoneme Frequency` (0.827). As can be seen from these loadings, the minimum frequency diphone is often the transitional diphone that connects the pronunciation of the first and the second character. This is the case for 76.40% of the two-character words in the CLD.

The main effect of the minimum diphone frequency of the word is presented in the left panel of Figure 3.32. Similar to the phoneme and diphone frequency measures discussed above, the confidence intervals are wide for predictor values near the edges of the predictor range. As before, pronunciation durations are longer for words with more frequent transitional diphones for non-extreme predictor values. Again, however, the effect size for the part of the predictor range where confidence intervals are narrow is small (10 ms).

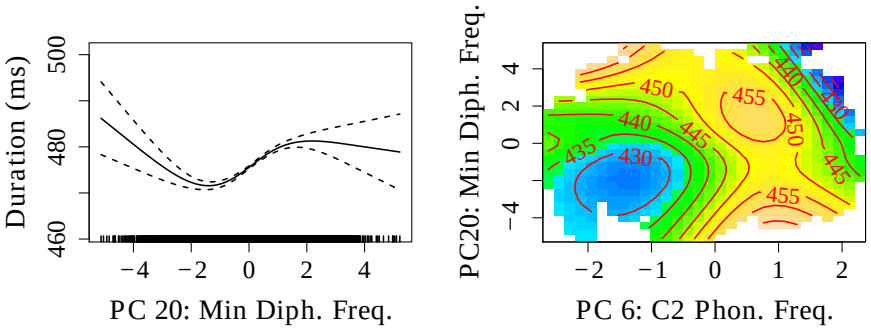

Figure 3.32: Duration results: two-character words. Minimum diphone frequency (left panel) and the interaction character 2 phoneme frequency with minimum diphone frequency (right panel). Right panel shows the additive contour surface for the main effect of character 2 phoneme frequency, the main effect of minimum diphone frequency and the interaction between character 2 phoneme frequency and minimum diphone frequency.

We furthermore observed an interaction between `PC6` (which, as a reminder, describes the phoneme frequency of the second character) and `PC20`. The interaction between `PC6` and `PC20` ($F = 9.017$, $p < 0.001$) is presented in the right panel of Figure 3.32, which shows that the main effect of the minimum diphone frequency of the word is present for words with a low average phoneme frequency for the second character only. For words with a high average phoneme frequency of the second character, the effect of minimum diphone frequency seems to reverse. However, the number of data points with high values for both predictors is limited. As a result, the decreased pronunciation durations in the upper right corner of the right panel of Figure 3.32 are not statistically robust.

Thus far we described frequency effects, both at a lexical and a sub-lexical level. The GAM for the pronunciation durations of one-character words furthermore revealed an effect of the number of phonemes, with pronunciation durations being 29 ms longer for high predictor values as compared to low predictor values. The GAM fitted to the pronunciation durations for two-character words likewise showed effects of the number of phonemes, both for the first and for the second character. The number of phonemes of the first character is encoded in `PC43` (lexical predictor with the highest loading: `Character 1 Phonemes` (0.659)), whereas the number of phonemes of the second character is encoded in `PC38` (highest loading: `Character 2 Phonemes` (0.734)).

The effects of PC43 ($F = 69.885$, $p < 0.001$) and PC38 ($F = 37.857$, $p < 0.001$) are presented in Figure 3.33. For both the first character (left panel) and the second character (right panel), pronunciation durations increase for characters with a greater number of phonemes. Whereas the effect of the number of phonemes was linear for one-character words, the effects for the number of phonemes in the first and second character of two-character words show some non-linearities. For the first character, the effect of the number of phonemes is most prominent for low-to-medium predictor values and levels off for high predictor values. The effect of the number of phonemes for the second character shows the opposite pattern of results, with an effect that is prominent for high predictor values and absent for low-to-medium predictor values.

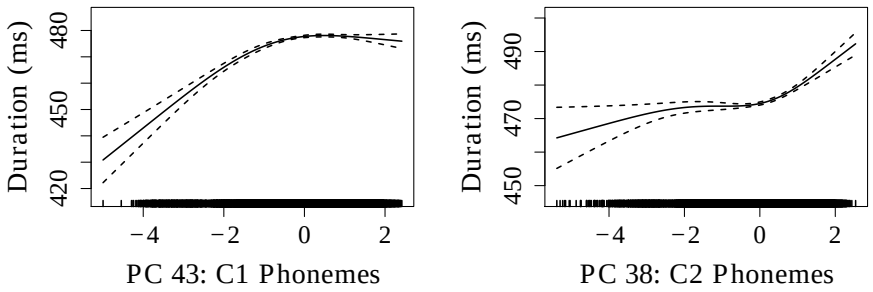

Figure 3.33: Duration results: two-character words. Number of phonemes for character 1 (left panel) and character 2 (right panel).

At 28 ms, the effect size of the number of phonemes for the second character is similar to the effect size of the effect of the number of phonemes for one-character words, which was 29 ms. For the first character, the effect of the number of phonemes is stronger, with an effect size of 47 ms. The greater effect size for the number of phonemes in the first character as compared to the number of phonemes in the second character fits well with the observation that the average pronunciation duration for two-character words (472 ms) is only 49.37% longer than the average pronunciation duration for one-character words (316 ms). This suggests that pronunciation durations for second characters are considerable shorter than pronunciation durations for first characters. As a result, the absolute effect size of the number of phonemes is smaller for two-character words than for one-character words.

Thus far, we have discussed the effects of phonological frequency and phonological complexity on the pronunciation durations for two-character words. A third phonological property that influenced pronunciation durations is phonological neighbourhood density. Figure 3.34 presents the effects of the principal components that encode the phonological neighbourhood density of the first and second character, `PC18` ($F = 56.981$, $p < 0.001$) and `PC19` ($F = 28.008$, $p < 0.001$). The loading on `PC18` is strongly positive for `Character 1 PLD` (0.926) and strongly negative for `Character 1 Phonological N` (-0.941). Similarly, `Character 2 PLD` has a strong positive loading on `PC19` (0.916), whereas `Character 2 Phonological N` has a strong negative loading (-0.945) on this principal component. Low predictor values for `PC18` and `PC19` thus reflect a high phonological neighbourhood density for the first and second character, respectively. The opposite loadings for the measures of phonological Levenshtein distance and phonological neighbourhood density are as expected: the more phonological neighbours a character has, the smaller the average distance between a character and its 20 closest phonological neighbours.

Figure 3.34 presents the effects of `PC18` (left panel) and `PC19` (right panel). The x-axes of both panels in Figure 3.34 are reversed for ease of interpretation (i.e., the right sides of both panels correspond to high values of neighbourhood density). As can be seen in the left panel of Figure 3.34, a high phonological neighbourhood density leads to shorter pronunciation durations for the first character. By contrast, the right panel of Figure 3.34 indicates that a greater number of phonological neighbours results in longer pronunciation durations for the second character. Both effects have relatively small effect sizes (effect size `PC18`: 14 ms; effect size `PC19`: 10 ms) and are most prominent for low predictor values (i.e., a high number of phonological neighbours).

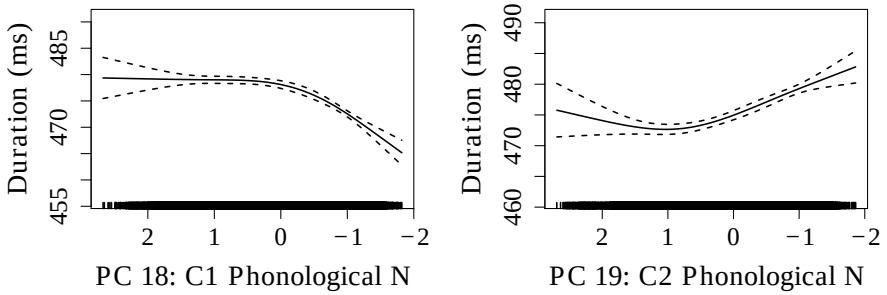

Figure 3.34: Duration results: two-character words. Phonological neighbourhood for character 1 (left panel) and character 2 (right panel). The x-axes are reversed for ease of interpretation.

The effect of phonological neighbourhood density for the first character is in line with findings by Gahl et al. (2012), who found increased phonetic reduction for words with large phonological neighbourhoods in English. `Phonological N` is the first predictor for which we clearly observed an opposite pattern of results for the first and second character. As we demonstrate in the next section, however, the reversal of a lexical predictor effect for the first and second character is ubiquitous in our analysis of the eye fixation durations.

The last two principal components that showed a significant effect on the pronunciation durations for two-character words describe the consistency of the mapping between orthography and phonology. First, consider the effect of `PC12` ($F = 12.934$, $p < 0.001$), which is depicted in Figure 3.35. As a reminder, `PC12` encodes the number of homographs and the frequency of these homographs for the first character. As can be seen in Figure 3.35, pronunciation durations are somewhat shorter when the first character has few pronunciations. However, with an effect size of a mere 8 ms, the effect of `PC12` is extremely subtle.

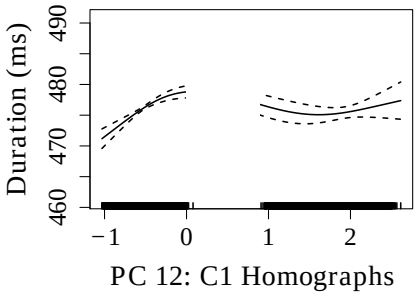

Figure 3.35: Duration results: two-character words. Character 1 homographs.

Second, we observed an effect of `PC9`. The lexical predictors with the highest loadings on `PC9` are `Character 2 Homophones (Tokens)` (0.950), `Character 2 Homophones Frequency` (0.933) and `Character 2 Homophones (Types)` (0.867). The main effect of `PC9` ($F = 23.847$, $p < 0.001$) is presented in the left panel of Figure 3.36. Pronunciation durations increase as the number of characters that have the same pronunciation increases. Predicted pronunciation durations for the lowest predictor values are 22 ms shorter than predicted pronunciation durations for the highest predictor values.

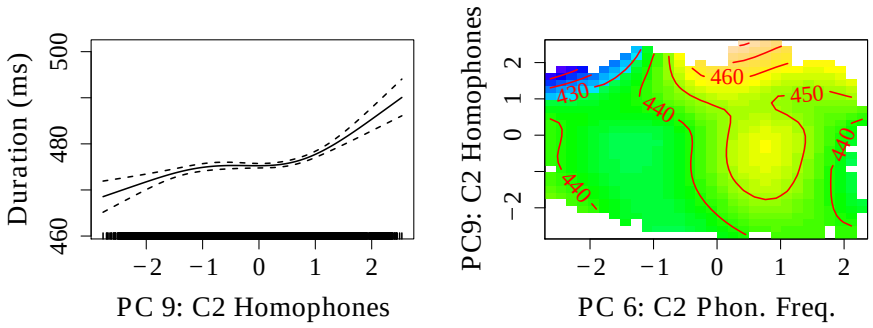

Figure 3.36: Duration results: two-character words. Character 2 homophones (left panel) and the interaction of character 2 homophones with character 2 phoneme frequency (right panel). Right panel shows the additive contour surface of the main effect of character 2 homophones, the main effect of character 2 phoneme frequency and the interaction between character 2 homophones and character 2 phoneme frequency.

`PC9`, however, interacts with `PC6` (i.e., the average phoneme frequency of the second character). The interaction between `PC6` and `PC9` ($F = 11.551$, $p < 0.001$) is presented in the right panel of Figure 3.36. For second characters with a high average phoneme frequency the homophone effect is as described above: a greater number of homophones results in longer pronunciation durations. Most likely, this effect is a result of decreased association strength between a lexical representation and the corresponding combination of phonological features when this combination of phonological features is shared across many characters. For second characters with low phoneme frequencies, however, the effect seems to reverse. It should be noted, however, that there are relatively few data points in the top left corner of the right panel of Figure 3.36. It is therefore not clear how robust the reversal of the effect of low values of `PC6` is.

In addition to the effects of the numerical predictors described above, we observed an effect of two categorical variables: `Character 1 Tone` and `Character 2 Tone`. The effect of `Character 1 Tone` is subtle. The only pairwise comparisons that reached significance at the Bonferroni-corrected α level were the pairwise comparisons between Tone 1 (reference level) on the one hand and Tone 3 (+4 ms) and Tone 4 (+3 ms) on the other hand.

Consistent with the GBM analysis, the effect of tone is larger for the second character than for the first character. Pronunciation durations were longest for Tone 2 (+18 ms), followed by Tone 1 (reference level), Tone 5 (−12 ms), Tone 4

(−18 ms) and Tone 3 (−20 ms). All pairwise comparisons were significant at the Bonferroni-corrected α level, with the exception of the comparisons between Tone 3 and Tone 4 and between Tone 4 and Tone 5.

Finally, a post-hoc analysis for the subset of the two-character words for which the first character contains a phonetic radical revealed an effect of `PC3P`, which describes the phonology-to-orthography consistency at both the character level and the level of the phonetic radical. The lexical predictors with the highest loadings on `PC3P` are `Character 1 PR Backward enemies (Tokens)` (loading: 0.943), `Character 1 PR Backward Enemies Frequency` (0.904), `Character 1 Backward Enemies (Types)` (0.891), `Character 1 Homophones (Types)` (0.897), `Character 1 Homophones (Tokens)` (0.894) and `Character 1 Homophones Frequency` (0.827). The effect of `PC3P` ($F = 18.998$, $p < 0.001$) is presented in Figure 3.37.

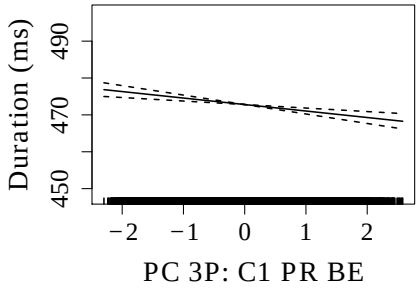

Figure 3.37: Duration results: two-character words (post-hoc analysis). Character 1 phonology-to-orthography consistency.

The effect is linear in nature, with pronunciation durations being shorter for characters with an inconsistent phonology-to-orthography mapping as compared to characters with a consistent phonology-to-orthography mapping. The effect of `PC3P` is in the opposite direction of the effect of the number of homophones for the second character reported above. However, the effect size of the effect of `PC3P` is a mere 9 ms. Even though the effect of `PC3P` is statistically significant, it is therefore not clear how meaningful this effect is.

A post-hoc analysis for the subset of the two-character words for which the second character contains a phonetic radical furthermore revealed an effect of the categorical variable `Character 2 PR Regularity` (t = -5.281, $p < 0.001$). Predicted pronunciation durations are 4 ms shorter for words in which the second character pronunciation is identical to the pronunciation of its phonetic radical. However,

given the small effect size of this effect, the influence of phonetic radical regularity on pronunciation durations for two-character words is limited.

3.4.3 Eye fixation durations

3.4.3.1 GBM

Table 3.3 presents the relative influences for the 156 predictor in the GBM fitted to the eye fixation durations. Underlining the importance of taking the time course of lexical processing into account – and carrying out separate GAM analyses for fixations that start at different points in time – the variable with the highest relative influence was `Fixation Start Time` (i.e., the time in ms at which a fixation starts, relative to stimulus onset). By itself, `Fixation Start Time` accounted for nearly half of the explanatory power of the GBM (relative influence: 47.435%).

We added two more control variables to the GBM fitted to the eye fixation durations in comparison to the GBMs reported earlier: `X Position` and `Y Position`. These control variables denote the horizontal and vertical position of a fixation on the screen. Both `X Position` (16.869%) and `Y Position` (1.377%) have substantial explanatory power for the fixation durations. This indicates that not only the time at which a fixations starts, but also the physical location of a fixation co-determines fixation durations.

As was the case for naming latencies and pronunciation durations, the GBM furthermore showed effects of `Session` (3.576%) and `Trial` (1.350%). This leads to a total relative influence of the control variables of 71.147%. For the investigation of eye fixation durations controlling for the influence of non-linguistic predictors thus is even more pivotal than it was for naming latencies and pronunciation durations.

Figure 3.38 provides an overview of the summed relative importance of the predictors in each cluster and group of predictors. Apart from the overwhelming influence of the control variables, the relative importance of predictors is more evenly spread out across the 6 groups of numerical predictors as compared to the relative importance of predictors for the naming latencies and pronunciation durations.

The group of predictors with the greatest summed relative influence is Group 1, which contains frequency-related predictors. The predictors in Group 1 have a summed relative influence of 11.076%. This summed relative importance is spread out across the six clusters in Group 1, with summed relative influences of 2.347% for Cluster 1 (word frequency), 2.347% for Cluster 2 (association measures), 1.749%

Table 3.3: Relative variable influences in an XGBoost model fitted to the fixation durations. Abbreviations: BE = Backward Enemies, C1 = Character 1, C2 = Character 2, Diph. = Diphone, Freq. = Frequency, HLC = High-Level Components, LLC = Low-Level Components, Phono = Phonological, Phon. = Phoneme, SUBTL = SUBTLEX-CH, Typ. = Types, Tok. = Tokens.

rank	predictor	infl.	rank	predictor	infl.
1	Fixation Start Time	47.435	29	C2 Friends	0.320
2	X	16.869	30	Mean Phon. Freq.	0.310
3	Session	3.576	31	Strokes	0.306
4	C1 Entropy	2.509	32	C1 Mean LLC Freq.	0.293
5	Y	1.377	33	PMI	0.281
6	Trial	1.350	34	Mean Diph. Freq.	0.280
7	C2 Structure	0.886	35	C1 Trigram Entropy	0.273
8	C2 Pixels	0.799	36	C2 SR Family Size	0.263
9	CD (Gigaword)	0.781	37	C2 Mean LLC Freq.	0.259
10	Entropy Char. Freqs.	0.751	38	C2 Friends Freq.	0.256
11	C2 Strokes	0.730	39	C2 Pixels OLD	0.256
12	C2 Family Size	0.670	40	Frequency (SCCoW)	0.255
13	C1 Pixels	0.610	41	C1 Min HLC Freq.	0.252
14	C1 Strokes	0.446	42	C2 LLC OLD	0.236
15	t-Score	0.408	43	CD (SCCoW)	0.227
16	C2 RE	0.399	44	C2 Min HLC Freq.	0.225
17	Frequency (Gigaword)	0.386	45	Initial Phoneme	0.217
18	C2 Phono N	0.377	46	Min Diph. Freq.	0.215
19	CD (SUBTL)	0.376	47	C2 Min Diph. Freq.	0.212
20	C2 Entropy	0.367	48	C1 Mean HLC Freq.	0.207
21	C2 Freq. (SUBTL)	0.356	49	C1 CD (SUBTL)	0.205
22	C1 Pixels OLD	0.343	50	C1 Homophones Freq.	0.204
23	C1 RE	0.337	51	Traditional Freq.	0.204
24	C1 Picture Size	0.336	52	C2 Min LLC Freq.	0.203
25	C2 Picture Size	0.328	53	C2 LLC	0.201
26	C2 Trigram Entropy	0.323	54	C1 Homophones (Tok.)	0.200
27	Frequency (SUBTL)	0.322	55	Transitional Diph. Freq.	0.199
28	Final Phoneme	0.321	56	C2 Mean HLC Freq.	0.199

Table 3.3 (continued)

rank	predictor	infl.	rank	predictor	infl
57	C1 Family Freq.	0.195	90	C1 Max HLC Freq.	0.117
58	C1 LLC OLD	0.194	91	C2 Initial Diph. Freq.	0.117
59	Position-specific PMI	0.194	92	C1 Freq. (SCCoW)	0.116
60	C2 Homophones Freq.	0.190	93	C2 Max Diph. Freq.	0.115
61	C1 Min LLC Freq.	0.187	94	C2 CD (Gigaword)	0.111
62	C2 PR Friends Freq.	0.183	95	C2 Traditional Freq.	0.108
63	C1 Min Diph. Freq.	0.182	96	C1 Traditional Freq.	0.107
64	C1 SR Frequency	0.180	97	C1 SR Strokes	0.105
65	C1 Freq. (SUBTL)	0.179	98	C1 PR Enemies Freq.	0.105
66	C2 Freq. (Gigaword)	0.177	99	C1 CD (Gigaword)	0.105
67	C1 Friends Freq.	0.175	100	C2 Homophones (Typ.)	0.100
68	C1 LLC	0.169	101	C1 Family Size	0.096
69	C1 Phono N	0.164	102	C2 Freq. (SCCoW)	0.094
70	C2 CD (SUBTL)	0.163	103	C2 Homographs Freq.	0.093
71	C1 PR Friends Freq.	0.162	104	C1 Friends	0.091
72	C1 CD (SCCoW)	0.159	105	C2 PLD	0.091
73	C2 Mean Diph. Freq.	0.151	106	Max Diph. Freq.	0.091
74	C1 PR Frequency	0.151	107	Phono N	0.090
75	C1 Mean Diph. Freq.	0.148	108	C2 Tone	0.090
76	C1 SR Family Size	0.143	109	C2 CD (SCCoW)	0.089
77	C1 Freq. (Gigaword)	0.143	110	C1 Homographs Freq.	0.083
78	C1 Structure	0.142	111	C1 Max Diph. Freq.	0.079
79	C1 PR BE Freq.	0.142	112	PLD	0.079
80	C2 PR BE Freq.	0.139	113	C1 PR Family Size	0.078
81	C1 HLC	0.135	114	C2 Max HLC Freq.	0.077
82	C2 Homophones (Tok.)	0.134	115	C2 PR Friends	0.077
83	C2 Mean Phon. Freq.	0.133	116	C1 PLD	0.075
84	C1 PR BE (Tok.)	0.131	117	C2 PR Enemies (Tok.)	0.073
85	Min Phon. Freq.	0.129	118	C2 Min Phon. Freq.	0.072
86	C1 Mean Phon. Freq.	0.126	119	C1 Min Phon. Freq.	0.071
87	C1 PR Enemies (Tok.)	0.124	120	C1 Initial Diph. Freq.	0.070
88	C1 PR Friends	0.123	121	C1 Homophones (Typ.)	0.069
89	C2 SR Frequency	0.118	122	C2 PR Frequency	0.066

Table 3.3 (continued)

rank	predictor	infl.	rank	predictor	infl.
123	C2 Family Freq.	0.063	140	C2 LLC N	0.036
124	C2 Initial Phon. Freq.	0.063	141	C2 PR Enemies (Typ.)	0.035
125	C2 PR Strokes	0.062	142	C2 Type	0.035
126	C1 PR Strokes	0.059	143	C2 PR BE (Typ.)	0.034
127	C2 PR Family Size	0.059	144	C2 Max LLC Freq.	0.032
128	C1 Max LLC Freq.	0.059	145	C2 HLC	0.032
129	C2 PR BE (Tok.)	0.059	146	Phonemes	0.032
130	C1 LLC N	0.057	147	C1 Max Phon. Freq.	0.031
131	C2 Phonemes	0.054	148	Max Phon. Freq.	0.028
132	C1 Initial Phon. Freq.	0.054	149	C1 Type	0.020
133	C2 PR Enemies Freq.	0.054	150	C1 Phonemes	0.019
134	C2 SR Strokes	0.054	151	C2 Max Phon. Freq.	0.018
135	C1 PR BE (Typ.)	0.053	152	C1 Homographs (Typ.)	0.016
136	C1 Tone	0.048	153	C2 Homographs (Typ.)	0.011
137	C1 Homograph (Tok.)	0.044	154	C1 PR Regularity	0.008
138	C2 Homograph (Tok.)	0.041	155	Length	0.002
139	C1 PR Enemies (Typ.)	0.040	156	C2 PR Regularity	0.000

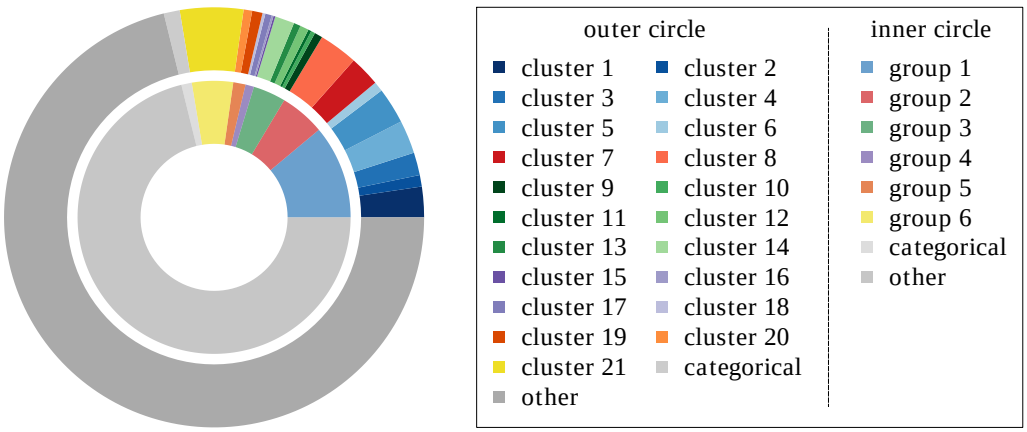

Figure 3.38: Relative variable influence per cluster in an XGBoost model fitted to the fixation durations.

for Cluster 3 (character 1 frequency), 2.559% for Cluster 4 (character 2 frequency), 2.782% for Cluster 5 (character 1 entropy) and 0.756% for Cluster 6 (character 2 entropy).

The frequency-related predictor that co-determines fixation durations to the greatest extent is `Character 1 Entropy`, which has a relative influence of 2.509%. With a relative influence of 0.367%, the entropy of a second character (`Character 2 Entropy`) is also among the 20 predictors with the highest relative influence. A third predictor in the group of frequency-related measures that had a relatively high relative influence is a measure that describes the association between the first and second character: `t-Score` (relative influence: 0.408%). The effects of these measures demonstrate that the information-theoretic properties of one- and two-character words provide importance guidance for the eye movement patterns for these words.

Furthermore, three measures of the frequency of the word as a whole are in the top 20 predictors with the highest relative influence: `CD (Gigaword)` (0.781%), `Frequency (Gigaword)` (0.386%) and `CD (SUBTLEX-CH)` (0.376%). For the eye fixation durations the frequency measures from the Gigaword corpus therefore provide more explanatory power as compared to the frequency measures from the SCCoW and SUBTLEX-CH.

Compared to the influence of word-level frequency measures, the influence of character-level frequency measures on the eye fixation durations is limited. No frequency measures of the first character were among the 20 predictors with the highest relative influence. For the second character, one predictor co-determined fixation durations to a considerable extent: `Character 2 Family Size` had a relative influence of 0.670%.

The group of numerical predictors with the second highest summed relative influence is Group 2. As a reminder, the numerical variables in Group 2 describe the visual complexity of the word and its characters. The summed relative influence of the predictors in Group 2 is 5.335%, with a somewhat greater summed influence for measures describing the visual complexity of the second character (Cluster 8; summed relative influence: 2.986%) as compared to measures describing the visual complexity of the first character (Cluster 7; summed relative influence: 2.349%).

For both the first and the second character, the pixel count in the corresponding image file is among the 20 predictors with the highest relative influence. The relative influence for `Character 1 Pixels` is 0.610%, whereas the relative influence for

`Character 2 Pixels` is 0.799%. Similarly, the number of strokes for both characters has an effect on the eye fixation durations, with relative influences of 0.446% for `Character 1 Strokes` and 0.730% for `Character 2 Strokes`.

The GBM for pronunciation durations was dominated by predictors from Group 3, which contains phonological measures. For naming latencies, the effects of phonological measures were more subtle. At 3.875%, the summed relative influence of the measures in Group 3 for the eye fixation durations is modest. The only phonological predictor among the 20 predictors with the highest relative influence is `Character 2 Phonological N` (relative influence: 0.377%).

The explanatory power of the variables in Group 4 (homographs; summed relative influence: 1.007%) and Group 5 (homophones; summed relative influence: 1.455%) is even more limited. The only other numerical predictors with considerable predictive power are two measures from Group 6 (other predictors; summed relative influence: 4.879%): `Entropy Character Frequencies` (0.751%) and `C2 RE` (0.399%). The relatively high variable importance of these predictors provides further support for the idea that the information-theoretic properties of a word co-determine eye movement patterns to a substantial degree. Finally, the top 20 predictors with the highest relative influence include a categorical variable: `Character 2 Structure` (0.886%).

The GBM analysis provides an overview of the overall influence of the predictors in the CLD on the eye movement durations. The GBM reported here, however, provides no insight into the time course of lexical processing. Furthermore, the qualitative nature of the effects of individual predictors remains unclear. Therefore, we now turn to an analysis of the eye movement patterns using GAMs.

3.4.3.2 GAM

3.4.3.2.1 One-character words

As noted above, we analyzed the eye fixation data through a series of analyses on moving time windows, in which data points eye fixations are grouped together on the basis of the point in time at which they started, relative to stimulus onset. We analyzed eye fixation durations for a total of 10 time windows, starting with fixations that started between −400 and −200 ms before stimulus onset and ending with fixations that started between 1400 and 1600 ms after stimulus onset.

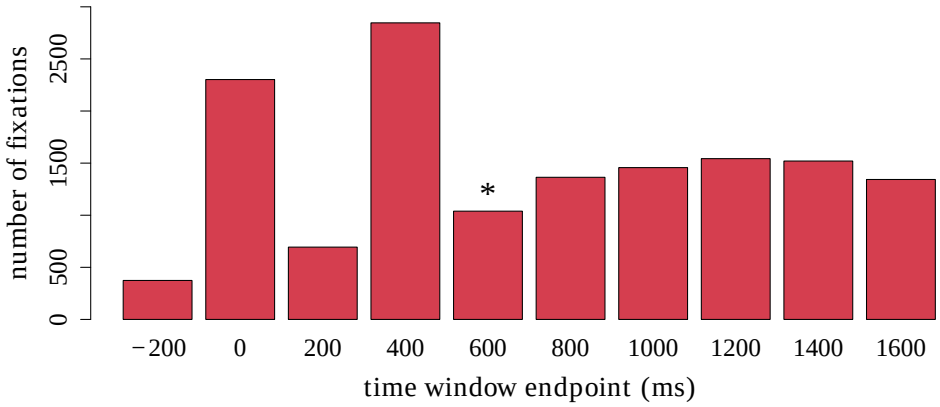

Figure 3.39: Number of fixations on one-character words for each time window. Times are the endpoints of the 200 time windows. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (76.30%) is marked by an asterisk.

The number of data points in each time window is shown in Figure 3.39, whereas the average duration of the fixations in each time window is shown in Figure 3.40. Times in Figure 3.39 and Figure 3.40 are the endpoints of each time window. For instance, the bar denoted with 200 shows the number of fixations that start in the 0 to 200 ms time window. We only included fixations that end after stimulus onset in our analyses. As a result, there were very few data points in the −400 to −200 ms time window (374) and the fixations in this time window have relatively long durations (mean fixation duration: 547.51 ms).

For the −200 to 0 ms time window 2,302 data points are available. The fixations starting in the −200 to 0 ms time window and ending after stimulus onset had an average duration of 417.77 ms. The average time at which the 2,302 fixations in the −200 to 0 ms time window ended was 281.42 ms after stimulus onset, whereas the average naming latency for one-character words was 542.93 ms. Hence, for a large number of one-character words, a substantial amount of lexical processing is done during fixations that started prior to stimulus onset. This highlights the benefit of adopting an analysis strategy for the eye fixation patterns using moving windows, as compared to a more traditional approach in which the duration of first, second, and further fixations following stimulus onset is analyzed.

The start of relatively long fixations prior to stimulus onset is reflected in the limited number of fixations that start between 0 and 200 ms after stimulus onset (694 ms). A second peak in the distribution of the number of fixations is reached in the

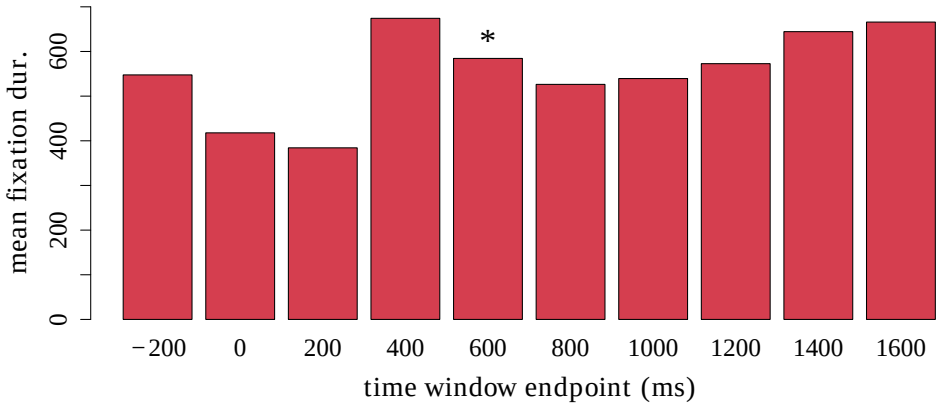

Figure 3.40: Average duration of fixations on one-character words for each time window. Times are the endpoints of the 200 time windows. Only fixations that end after stimulus onset are included. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (76.30%) is marked by an asterisk.

200 to 400 ms time window, for which no less than 2,846 data points are available. The average duration of fixations starting in this time window was 674.21 ms, which indicates that fixations starting in the 200 to 400 ms time window typically are the last fixations starting before the onset of the pronunciation.

The average number of fixations starting between stimulus onset and pronunciation onset was no more than 1.21 (0 fixations: 11.95%; 1 fixation: 61.38%, 2 fixations: 21.59%, 3 or more fixations: 5.07%). The distribution of fixation indices relative to stimulus onset over the time windows under investigation is presented in Figure 3.41. By definition, only fixations that started prior to stimulus onset exist for the −400 to −200 and −200 to 0 ms time windows. Fixations in the 0 to 200 ms time window were almost exclusively initial fixations following stimulus onset (98.13%). Even in the 200 to 400 ms time window, a large majority of fixations were first fixations relative to stimulus onset (84.89%). We therefore conclude that a single fixation after stimulus onset typically suffices to pronounce a one-character word.

Eye fixation patterns continue after the onset of pronunciation. The average number of fixations in the 400 to 600 ms through 1400 to 1600 ms time windows is 1,378. As can be seen in Figure 3.40, the average fixation duration for these later time windows is relatively long and gradually increases as a function of time from the 600 to 800 ms onwards. The fact that the participant continues to fixate on the word after pronunciation (onset) suggests that a substantial amount of post-

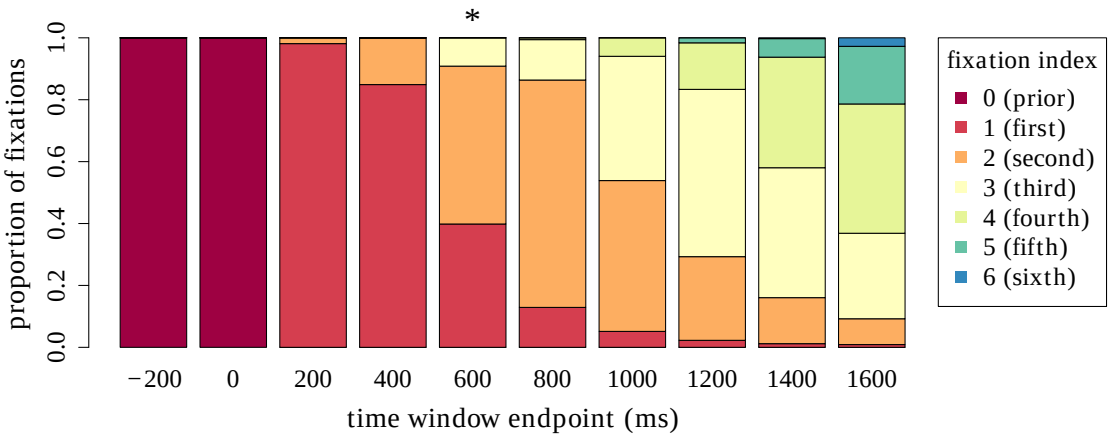

Figure 3.41: Proportion of fixation indices relative to stimulus onset for each time window. Times are the endpoints of the 200 time windows. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (76.30%) is marked by an asterisk.

processing takes places – potentially due to the nature of the experimental task (i.e., the word remained on the screen for 2000 ms).

Table 3.4 presents an overview of the predictor effects in the GAMs fitted to the durations of fixations starting in each time window. As before, the times in Table 3.4 are the endpoints of the 200 time windows. The column 200, for instance, shows the results of a GAM fitted to the fixations that start between 0 and 200 ms after stimulus onset. For each of the predictors that reached significance in at least one time window, we visually present the effect for a representative time window below.

Table 3.4: Overview of predictor effects on fixation durations for one-character words. Times are endpoints of 200 ms time windows. Plus symbols indicate a positive relation between predictor and dependent variable, minus symbols indicate a negative relation. Inverse U-shaped effects are indicated with the ∩ symbol.

	-200	0	200	400	600	800	1000	1200	1400	1600
Session		∩								
Trial				−						
X Position	∩	∩	∩	∩	∩	∩	∩	∩	∩	∩
PC1: C1 Frequency				+		+	+	+	+	
PC2: C1 Complexity				−						

We found a significant effect of `Session` ($F = 12.644$, $p < 0.001$) for fixations that start between −200 and 0 ms after stimulus onset. As can be seen in the left panel of Figure 3.42, the confidence intervals for the effect of `Session` are relatively wide. As a result, the exact nature of the effect of `Session` near the edges of the predictor range remains uncertain. Nonetheless, a clear decrease in fixation durations is visible between experimental sessions 10 and 20, with a difference of 51 ms between the longest and shortest predicted fixation durations.

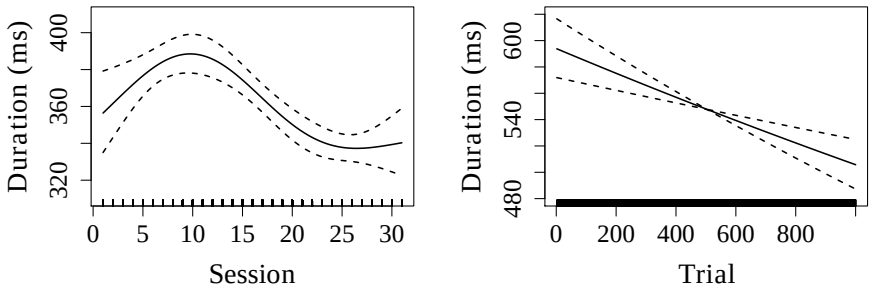

Figure 3.42: Fixation duration results: one-character words. Session for fixations that start between -200 and 0 ms after stimulus onset (left panel) and trial for fixations that start between 200 and 400 ms after stimulus onset (right panel).

In addition to the effect of `Session`, we found an effect of `Trial` for fixations that start between 200 and 400 ms after stimulus onset. The effect of `Trial` ($F = 18.397$, $p < 0.001$) is presented in the right panel of Figure 3.42. Fixation durations linearly decrease as a function of `Trial`, with a predicted difference of 88 ms between fixation durations for the first trial and fixation durations for the last trial.

The effects of `Session` and `Trial` only reached significance at a Bonferroni-corrected α level for a single time window. By contrast, the effect of a third control variable, `X Position`, reached significance in all time windows. The effect of `X Position` for the 200 to 400 ms time window ($F = 17.651$, $p < 0.001$) is presented in Figure 3.43.

Fixation durations are longest for fixations at a horizontal position of 840 pixels from the left edge of the screen. Given that we presented stimuli in the center of a screen with a horizontal resolution of 1680 pixels, this fixation position corresponds to the center of the word. Therefore, fixation durations are longest when the fixation position near the center of the word. When the fixation position deviates from the center of the word, fixations durations are almost 50% shorter (range of predicted values: 296 ms to 582 ms; effect size: 286 ms).

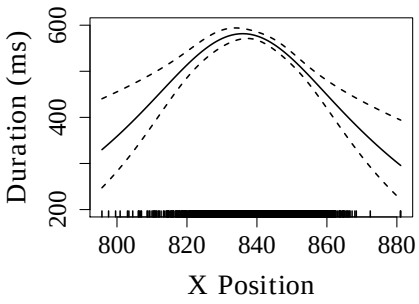

Figure 3.43: Fixation duration results: one-character words. Horizontal fixation position for fixations that start between 200 and 400 ms after stimulus onset.

In addition to the effects of the control variables `Session`, `Trial` and `X Position`, we found effects of two principal components: `PC1` and `PC2`. As a reminder, for one-character words `PC1` encodes the frequency of the first (and only) character, whereas `PC2` encodes its visual complexity. The effect of `PC1` reached significance in the 200 to 400, 600 to 800, 800 to 1000, 1000 to 1200 and 1200 to 1400 ms time windows and is presented for the 200 to 400 ms time window ($F = 32.994$, $p < 0.001$) in the left panel of Figure 3.44. The effect of `PC2` (see right panel of Figure 3.44) is more transient and is significant in the 200 to 400 ms time window only ($F = 31.805$, $p < 0.001$).

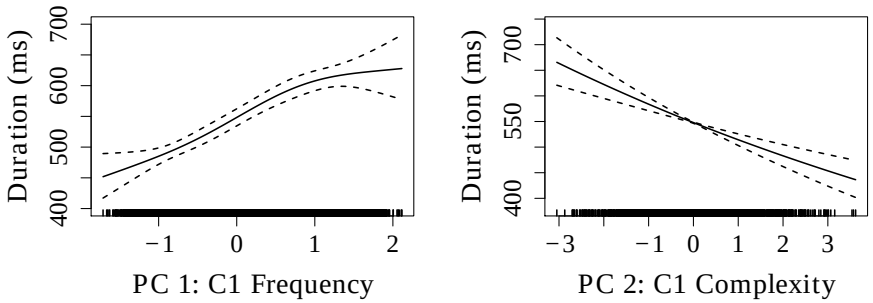

Figure 3.44: Fixation duration results: one-character words. Character frequency (left panel) and complexity (right panel) for fixations that start between 200 and 400 ms after stimulus onset.

As can be seen in Figure 3.44, the effects `PC1` and `PC2` are linear or near-linear. At 176 and 229 ms, respectively, the effect sizes of the effects of both `PC1` and `PC2` in the 200 to 400 ms time window are large. Interestingly, fixation durations are longer for characters with a high frequency and a low visual complexity as compared

to characters with a low frequency and a high visual complexity. Therefore, at least for one-character words, fixation durations are longer for words that are responded to faster. Rather than being indicative of additional processing, longer fixation durations may thus reflect a decreased need for refixations.

3.4.3.2.2 Two-character words

The number of fixations on two-character words for each time window is presented in Figure 3.45. The distribution of fixations per time window resembles the distribution of fixations per time window for one-character words. Again, there were relatively few fixations $(2,145)$ that started in the -400 to -200 ms time window that end after stimulus onset. By contrast, a large number of fixations $(14,817)$ started in the -200 to 0 ms time window end after stimulus onset. As can be seen in Figure 3.46, the average fixation duration of these fixations was nearly 400 ms (385.89 ms). As a result, the average endpoint of fixations that started in the -200 to 0 ms time window was 287.43 ms after stimulus onset. Therefore, as was the case for one-character words, substantial lexical processing takes place during fixations that started before stimulus onset.

Consistent with the distribution of fixations over time windows for one-character words, relatively few fixations $(9,645)$ start in the 0 to 200 ms time window, whereas more data points are available for the 200 to 400 $(12,626)$ and 400 to 600 $(14,078)$ ms time windows. The average naming latency for two-character words is 503.49

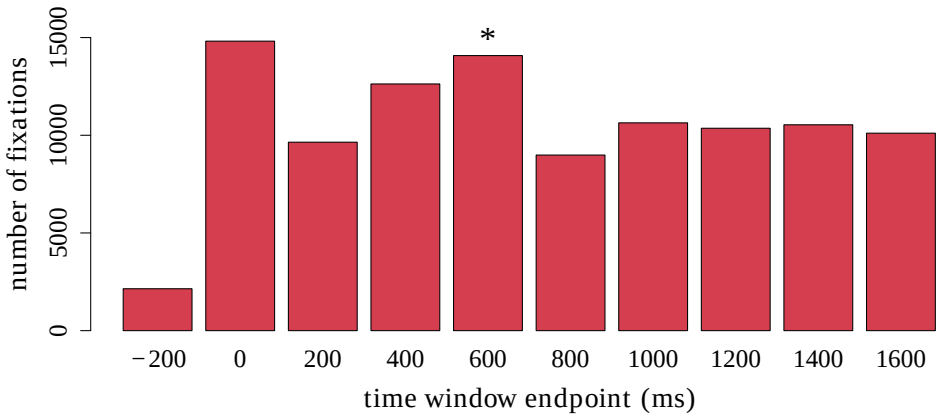

Figure 3.45: Number of fixations on two-character words for each time window. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (87.08%) is marked by an asterisk.

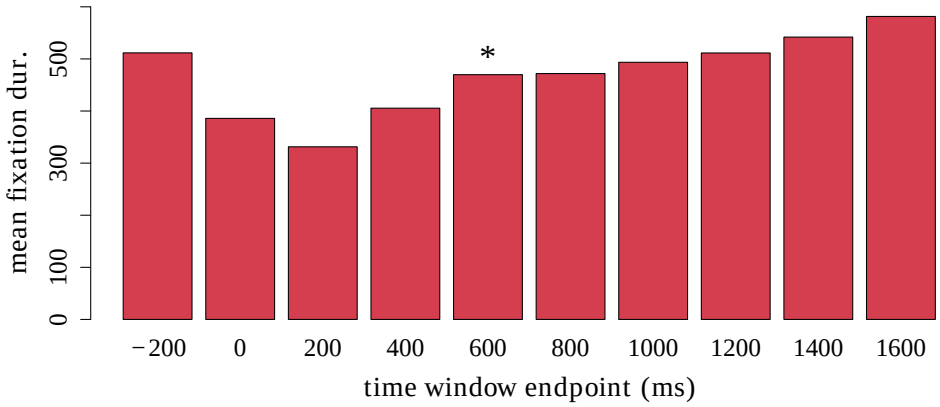

Figure 3.46: Average duration of fixations on two-character words for each time window. Only fixations that end after stimulus onset are included. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (87.08%) is marked by an asterisk.

ms. Nonetheless, the participant continued to fixate in the 600 to 800 through 1400 to 1600 ms time windows, with an average number of 10,784 fixations per time window. For fixations that started after stimulus onset, average fixation durations increase as a function of time (see Figure 3.46).

On average, the number of fixations starting between stimulus onset and pronunciation onset was 1.31 (0 fixations: 8.93%; 1 fixation: 54.29%, 2 fixations: 33.77%, 3 or more fixations: 3.00%). Two-character words therefore require slightly more fixations than one-character words, for which the average number of fixations starting between stimulus onset and pronunciation onset was 1.21. The proportion of fixation indices relative to stimulus onset for each time window is presented in Figure 3.47. Most fixations that start in the 0 to 200 (98.82%) and 200 to 400 (79,59%) ms time windows are the first fixation after stimulus onset. Fixations that start in the 400 to 600 ms time window typically are second fixations after stimulus onset (69.92% second fixations, 24.78% first fixations).

For one-character words, the horizontal fixation position showed a significant main effect on fixation durations in each time window, with longer fixation durations for fixations that were near the middle of the character. The effect of horizontal fixation position was orthogonal to the effects of lexical predictors. By contrast, as we document below, the horizontal position of a fixation determined the qualitative nature of the effects of a large number of lexical predictors for two-character words. Therefore, before discussing the results of the GAM analyses for two-character words, we provide some information about the development of fixation positions over time.

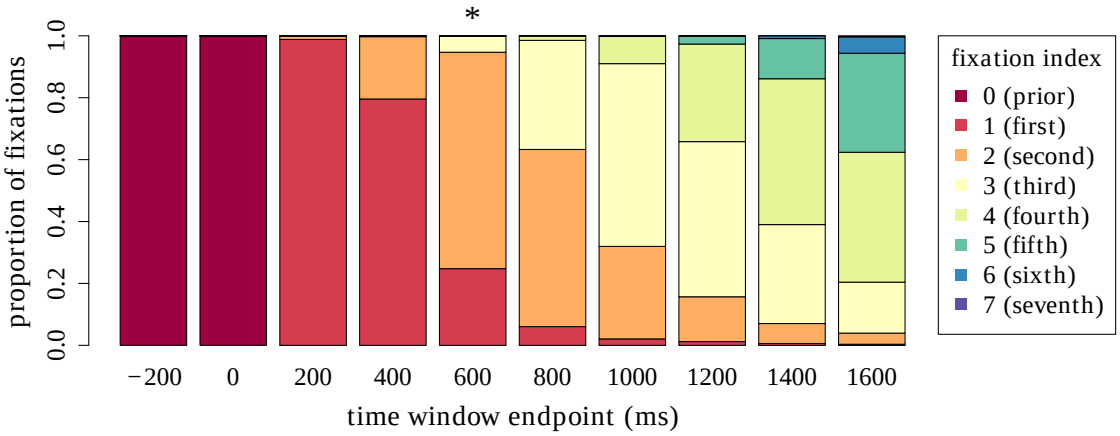

Figure 3.47: Proportion of fixation indices relative to stimulus onset for each time window. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (87.08%) is marked by an asterisk.

First, consider Figure 3.48, which shows the proportion of fixations on the first character (red bars) and the second character (blue bars) for each time window. Initially, most fixations are on the first character. The proportion of fixations on the first character is high in the −200 to 0 ms time window (83, 23%) and reaches a peak in the 0 to 200 ms time window. In this time window, no less than 98.12% of all fixations were on the first character. Later, the proportion of fixations on the second character increases. Fixations starting between 200 and 400 ms after

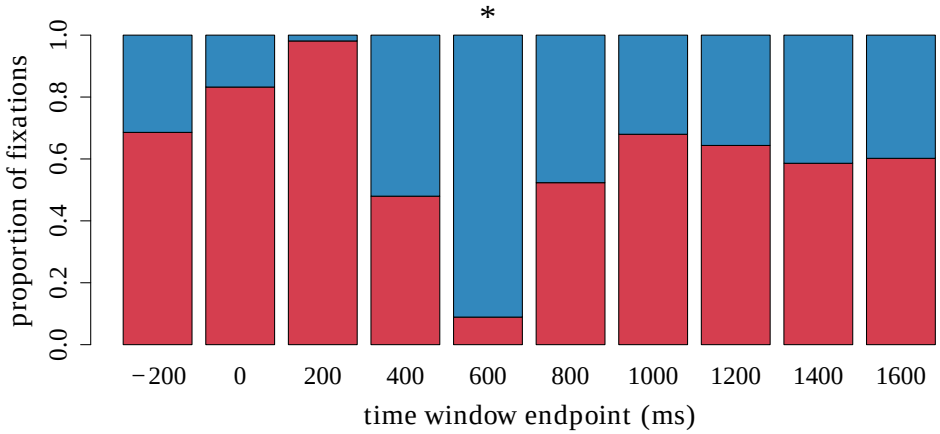

Figure 3.48: Proportion of fixations on the first (red bars) and second character (blue bars) of two-character words for each time window. The time window (400 to 600 ms after stimulus onset) in which most pronunciations start (87.08%) is marked by an asterisk.

stimulus onset are split evenly between the first (47, 97%) and the second (52.03%) character, whereas in 400 to 600 ms time window, 91.09% of all fixations are on the second character. During the early stages of lexical processing the typical reading pattern of our participant for two-character words thus was an initial fixation on the left character, followed by a fixation on the right character. After the 400 to 600 ms time window, the distribution of horizontal fixation positions stabilizes, with proportions of fixations on the first character ranging from 52.33% to 67.97% for fixations starting in the time windows between 600 and 1800 ms after stimulus onset.

Figure 3.49 further illustrates the development of fixation positions over time through scatterplots of the fixation position of fixations that started prior to stimulus onset and first fixations after stimulus onset (left panel) and the fixation position of first and second fixations after stimulus onset (right panel). As can be seen in the left panel of Figure 3.49, most fixations that started before stimulus onset were, unsurprisingly, near the center of the word. For these fixations, the first fixation after stimulus onset was typically on the left character. However, despite the fact that both one-character words and two-character words were presented in the experiment, the participant sometimes anticipated the potential presentation of a two-character word by fixating a little to the left of the center of the word. Whenever this was the case, first fixations after stimulus onset were often on the right character.

As noted above, the average number of fixations between stimulus onset and pronunciation onset was 1.31. As can be seen in the right panel of Figure 3.49, second fixations were almost never necessary when the first fixation after stimulus

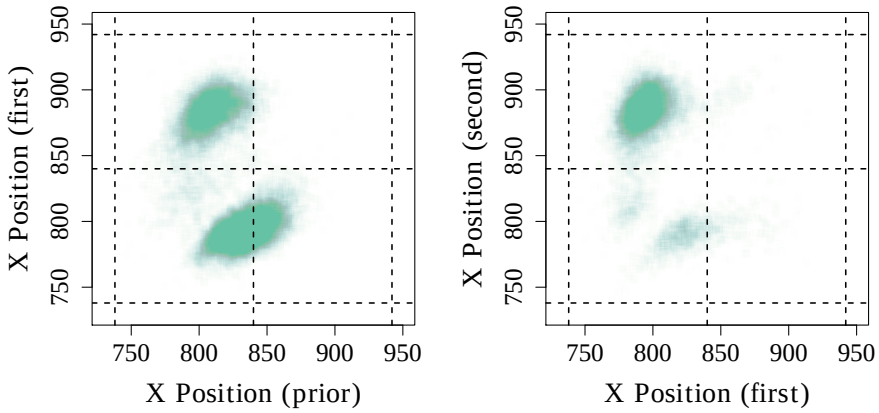

Figure 3.49: Horizontal fixation positions of prior and first fixations (left panel) and first and second fixations (right panel). Dotted lines indicate character borders.

onset was on the right character. Whenever a first fixation on the left character proved insufficient for pronunciation, second fixations were typically on the right character. Generally speaking, therefore, one fixation on the left character and one fixation on the right character were sufficient to pronounce a two-character word.

Perhaps surprisingly, average naming latencies are shorter for two-character words (503.49 ms) as compared to one-character words (542.93 ms). Below, we argue that the remarkable efficiency with which two-character words are processed is a result of a highly efficient processing strategy based on a dynamic search for information. Table 3.5 presents an overview of the effects of the lexical predictors and control variables on the fixation durations for two-character words. At the Bonferroni-corrected α level, we found effects of 5 control variables, 14 principal components and 3 categorical predictors. Table 3.5 indicates whether or not there was an effect of a predictor for each time window, but does not specify whether the predictor effect was a main effect, an interaction or a combination of a main effect and one or more interactions. More details about each effect will be provided in our discussion of the results below.

The first control variable that showed a significant effect is `Session`. We found a main effect of `Session` for all time windows, with the exception of the -400 to -200 ms time window. The effect of `Session` is presented for the 400 to 600 ms time window ($F = 52.451$, $p < 0.001$) in the left panel of Figure 3.50. For one-character words, the effect of `Session` was inverse U-shaped. Given the width of the confidence intervals, however, we could only be certain about the second part of the inverse U: a decrease in fixation durations over the course of the second half of the experiment. For two-character words, we see a similar inverse U-shaped effect, with an effect size of 65 ms. However, the confidence intervals for the effect of `Session` for two-character words are much less wide than the confidence intervals for the effect of `Session` for one-character words. For two-character words, therefore, we conclude that fixation durations increase over the course of the first half of the experiment and decrease over the course of the second half of the experiment.

Consistent with the findings for one-character words, we furthermore observed an effect of `Trial`. The main effect of `Trial` reached significance in the 0 to 200 and 200 to 400 ms time windows and is presented for the 0 to 200 ms time window ($F = 10.569$, $p < 0.001$; effect size: 21 ms) in the right panel of Figure 3.51. The effect of `Trial` is qualitatively similar to the effect of `Session`. Fixation durations increase during the first half of an experimental session and decrease during the second half of a session.

Table 3.5: Overview of predictor effects on fixation durations for two-character words. Time points are endpoints of 200 ms time windows (e.g., the 200 column shows results for fixations that start between 0 and 200 ms after stimulus onset). Plus symbols indicate a positive relation between predictor and dependent variable, minus symbols indicate a negative relation. Inverse U-shaped effects are denoted with the ∩ symbol, significant effects of categorical predictors are by stars and null effects by zeroes. The symbol C indicates a complex non-linear relation. The notation symbol1|symbol2 denotes an interaction between a lexical predictor and X Position, with symbol1 referring to the effect for fixations on the first character and symbol2 referring to the effect for fixations on the second character. All other interactions are omitted from this table.

	-200	0	200	400	600	800	1000	1200	1400	1600
Session		−	+	∩	∩	∩	∩	∩	∩	∩
Trial			∩	−						
Y Position		∩			−					
X Position	∩	∩	∩	C	C	−	−	∩	∩	∩
Final Phoneme					*					
PC1: C1 Frequency		−|+	−|+	−|+	+|+	+			+	+
PC2: C2 Frequency		+|0	+	+|−	+|−	+	+			
PC3: C2 Complexity		−|0	−|+	−|+	−|+		−			
PC4: C1 Complexity	+|−	+|−	+|−	+|−	−		+|−		−	−
PC5: Frequency				0|+		+	+	+	+	
PC22: C1 SR Freq		+			−|−					
PC23: C2 SR Freq			−	−|+						
PC8: C2 Diph. Freq.					−					
PC10: C1 Homoph.			+							
PC13: C2 Homogr.					C|0					
PC42: C1 Entropy			−	−|+						
PC39: C2 Entropy			+							
PC32: C2 RE				−						
PC36: H Char Freqs.		+			−|−					
C1 Type			*							
C2 Type			*	*						
C2 Tone					*					

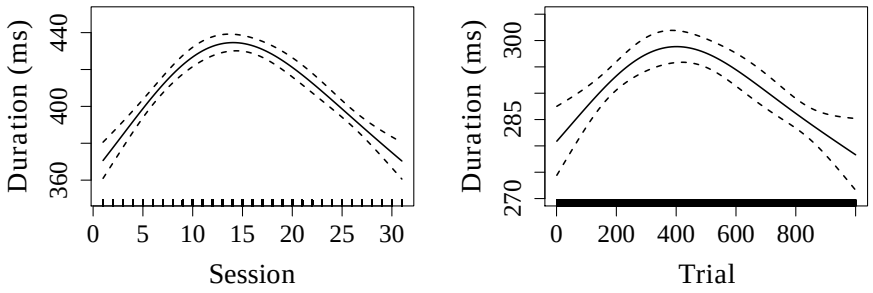

Figure 3.50: Fixation duration results: two-character words. Session for fixations that start between 0 and 200 ms after stimulus onset (left panel) and Trial for fixations that start between 400 and 600 ms after stimulus onset (right panel).

For the −200 to 0 ms time window we furthermore observed an interaction between `Session` and `Trial` ($F = 19.315$, $p < 0.001$). As can be seen in Figure 3.51, fixation durations are short for early trials in the first experimental sessions and late trials in the last experimental sessions. Therefore, during these parts of the experiment, less lexical processing occurred during fixations that started before stimulus onset.

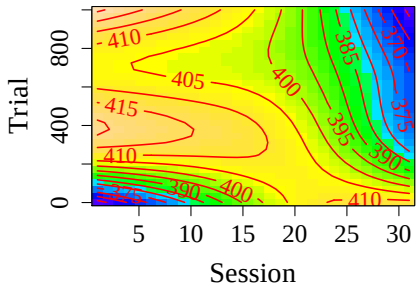

Figure 3.51: Fixation duration results: two-character words. Interaction between Session and Trial for fixations that start between -200 and 0 ms after stimulus onset. Additive contour surface of the main effect of Session, the main effect of Trial and the interaction between Session and Trial.

As was the case for the fixation durations for one-character words, we furthermore found a main effect of `X Position`, which was significant for all time windows. Figure 3.52 shows the effects of `X Position` for fixations that start between 0 and 200 ms after stimulus onset (left panel, $F = 193.192$, $p < 0.001$, effect size: 225 ms) and for fixations that start between 200 and 400 ms after stimulus onset (right panel, $F = 822.123$, $p < 0.001$, effect size: 183 ms). The dashed lines in Figure 3.52

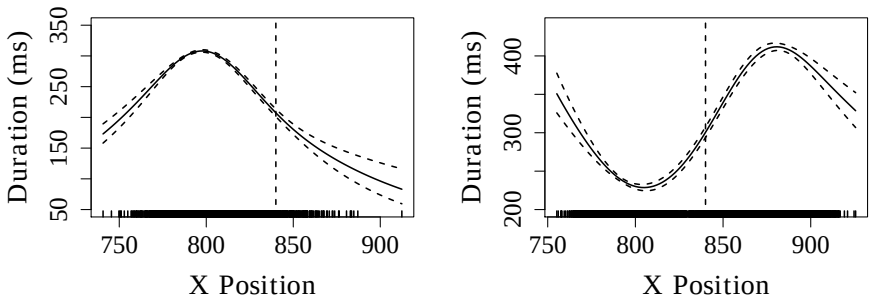

Figure 3.52: Fixation duration results: two-character words. Horizontal fixation position for fixations that start between 0 and 200 ms (left panel) and between 200 and 400 ms after stimulus onset (right panel). Dashed lines indicate the middle of the word.

indicate the middle of the word. Fixations on the left of the dashed line thus are on the left character, whereas fixations on the right of the dashed line are on the right character.

For one-character words, we observed increased fixation durations for fixations near the middle of the character. As can be seen in the left panel of Figure 3.52, fixation durations for fixations that start between 0 and 200 ms after stimulus onset are longest approaching the middle of the first character ($F = 193.192$, $p < 0.001$, effect size: 225 ms), which is at 789 pixels from the left border of the screen. By contrast, the right panel of Figure 3.52 shows that in 200 to 400 ms time window ($F = 822.123$, $p < 0.001$, effect size: 183 ms), fixation durations are longest for fixations approaching the middle of the second character, which is at 891 pixels from the left border of the screen. Furthermore, fixation durations in the 200 to 400 ms time window are particularly short for fixations that are near the middle of the first character. This pattern of results continues into the 400 to 600 ms time window. The effects of `X Position` indicate that the allocation of resources varies as a function of the processing demands at different points in time. At the earliest stages of lexical processing fixation durations are longer for fixations on the first character, whereas during later stages of lexical processing fixation durations are longer for fixations on the second character.

In addition to the main effect of `X Position`, we found an interaction of `X Position` with `Session` for the −200 to 0, 200 to 400, 1200 to 1400, 1400 to 1600 and 1600 to 1800 ms time windows. The interaction between `X Position` and `Session` for the 200 to 400 ms time window ($F = 7.111$, $p < 0.001$) is presented in Figure 3.53.

The effect size of `X Position` is large. Contour surfaces for interactions that include the main effect for `X Position`, therefore, are dominated by the effect of `X Position` to such an extent that a full appreciation of the main effect of the other predictor and the interaction between `X Position` and the other predictor becomes difficult. To allow for an easier interpretation of the interaction and the main effect of the predictor that interacts with `X Position`, all figures for an interaction with `X Position` exclude the main effect of `X Position` and include the additive contour surface of the interaction and the main effect of the predictor that interacts with `X Position` only. Figure 3.53, therefore, shows the additive contour surface for the main effect of `Session` and the interaction between `Session` and `X Position`.

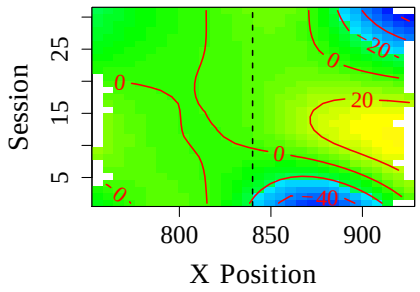

Figure 3.53: Fixation duration results: two-character words. Interaction between horizontal fixation position and session for fixations that start between 200 and 400 ms after stimulus onset. Figure shows the additive contour plot for the main effect of session and the interaction between session and horizontal fixation position.

As can be seen in Figure 3.53, the inverse U-shaped effect of `Session` is present for fixations on the second character only. Fixation patterns on the first character thus remain fairly constant throughout the experiment. This finding fits well with the fact that the effect of `Trial` was relatively subtle for one-character words, and was significant in the 0 to 200 ms time window only.

Not only the horizontal fixation position, but also the vertical fixation position co-determines fixation durations. The main effect of `Y Position` is presented in the left panel of Figure 3.54 for fixations that start between 400 and 600 ms after stimulus onset ($F = 16.551$, $p < 0.001$), but also reached significance in the -200 to 0 ms time window. The x-axis of the left panel Figure 3.54 is reversed for ease of interpretation. Fixations that are higher on the screen have longer durations as compared to fixations that are lower on the screen. The effect is linear and has an effect size of 91 ms.

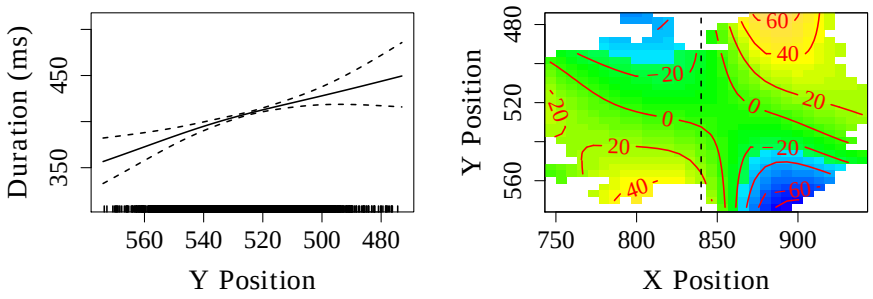

Figure 3.54: Fixation duration results: two-character words. Main effect of vertical fixation position (left panel) and the interaction of vertical fixation position with horizontal fixation position (right panel) for fixations that start between 400 and 600 ms after stimulus onset. The axes for vertical fixation position are reversed for ease of interpretation. The right panel shows the additive contour plot for the main effect of vertical fixation position and the interaction between vertical fixation position and horizontal fixation position.

The GAM analysis furthermore revealed an interaction between `X Position` and `Y Position`. The right panel of Figure 3.54 presents this interaction for the 400 to 600 ms time window ($F = 6.257$, $p < 0.001$). The y-axis of this Figure is reversed for ease of interpretation. The interaction is also significant in the −200 to 0 time window. As can be seen in the right panel of Figure 3.54, the main effect of `Y Position` is present for the 91.09% of the fixations in the 400 to 600 ms time window that are on the second character only. For the 8.91% of the fixations that are on the first character the effect reverses, with longer fixation durations for fixations near the top of the character. As will become clear below, this opposite pattern of results for fixations on the first and on the second character is a common pattern in the current data.

The fifth and final control variable that showed a significant effect on fixation durations is `Final Phoneme`. The effect of `Final Phoneme` reached significance in the 400 to 600 ms time window only ($F = 12.460$, $p < 0.001$). Fixation durations were particularly long for the nasal consonants "[n]" and "[ŋ]". The fact that `Final Phoneme` reaches significance in this time window is not surprising, given that 87.08% of all naming latencies are between 400 and 600 ms.

Consistent with the GAMs fitted to one-character words, we observed significant effects of the frequency of both the first (as encoded in `PC1`) and the second character (as encoded in `PC2`) on the fixation durations for two-character words. Consistent with the findings of G. Yan et al. (2006) character frequency thus influences fixation

durations on both characters in two-character words. The main effect of PC1 reached significance in the 0 to 200, 600 to 800, 1200 to 1400 and 1400 to 1600 time windows. The main effect of PC2 was significant in the −200 to 0, 0 to 200, 200 to 400, 600 to 800 and 800 to 1000 ms time windows. Upon closer inspection, however, it turned out that the effects of both PC1 and PC2 are best understood when their interactions with X Position are taken into account.

Figure 3.55 presents the interaction of PC1 (left panel; $F = 42.349$, $p < 0.001$) and PC2 (right panel; $F = 31.669$, $p < 0.001$) with X Position for fixations that start between 200 and 400 ms after stimulus onset. Again, the main effect of X Position (see Figure 3.52) is omitted from the contour plots in Figure 3.55 for ease of interpretation. The interactions between PC1 and X Position and between PC2 and X Position also reached significance in the −200 to 0 and 400 to 600 ms time windows. We furthermore observed an interaction between PC1 and X Position in the 0 to 200 ms time window.

First, consider the interaction between PC1 and X Position in the left panel of Figure 3.55. For fixations on the first character, a greater frequency leads to shorter fixation durations. In addition, we observed an effect of the frequency of the first character for fixations on the second character. The effect of the frequency of the first character for fixations on the second character is in the opposite direction of the effect of the frequency of the first character for fixations on the first character itself: fixation durations on the second character are longer when the first character has a higher frequency.[2]

The effect of first character frequency on fixation durations of fixations on the second character is possible due to a phenomenon referred to as parafoveal preview (see, e.g., Rayner et al., 1982; Rayner & Pollatsek, 1989). The parafovea is the region of the retina that surrounds the fovea. When fixating on a word, the fovea (i.e., the central part of the retina) processes the information in the immediate vicinity of the fixation position. Additional information about a broader visual area becomes available to the parafovea. This allows lexical elements that are not in the center of visual attention to be processed. A number of studies demonstrated that parafoveal preview effects are not limited to alphabetical languages, but occur in Mandarin Chinese as well (see, e.g., Inhoff & Liu, 1997; Inhoff, 1999; W. Liu et al.,

[2]For extreme values of X Position, we see additional reversals of the pattern of results. However, when limiting tensor product interactions to 4th order non-linearities in both dimensions, GAMs can be unreliable near the edges. Hence, it is unclear how statistically robust these additional reversals are.

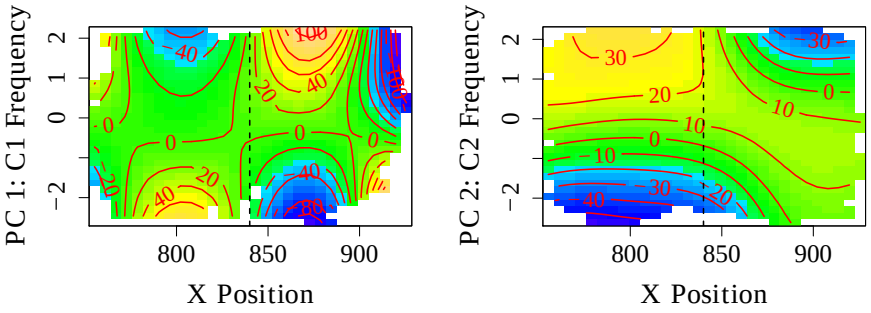

Figure 3.55: Fixation duration results: two-character words. Interaction with horizontal fixation position for character 1 frequency (left panel) and for character 2 frequency for fixations that start between 200 and 400 ms after stimulus onset (right panel). The main effect of horizontal fixation position is omitted from both contour plots for ease of interpretation.

2002; J. Yang et al., 2009; Tsai et al., 2004; M. Yan et al., 2009, 2012; Tsai et al., 2012).

The opposite effect of first character frequency for fixations on the first and second character is perhaps best understood as a quest for information. The greater the frequency of the first character, the less information it provides and the less long it needs to be fixated on to reach an understanding of the first character that is sufficient for an accurate response. Therefore, when fixations are on the first character, a short fixation suffices before attention is directed to the second character. By contrast, for fixations on the second character, there is little incentive to move back to the uninformative first character. Instead, all attention can be directed at the second character, which leads to longer fixation durations.

As can be seen in the right panel of Figure 3.55, the interaction between `PC2` and `X Position` is consistent with such an interpretation. Roughly speaking, the pattern of results for the frequency of the second character is the opposite of the pattern of results for the frequency of the first character: fixations durations on the first character increase as a function of the second character frequency, whereas fixation durations on the second character decrease. Again, this suggests that the presence of a highly frequent character results in increased attention for the other character.

The effect of both the first and the second character frequency is largest when the fixation is on the other character. The effect size for the effect of `PC1` in the 200 to 400 ms time window is 145 ms for fixations on the first character and 266 ms for

fixations on the second character. Similarly, the effect size for the effect of PC2 is 141 ms for fixations on the second character and 174 ms for fixations on the first character.

The interactions of PC1 and PC2 with X Position highlight two important characteristics of lexical processing for two-character words that are pivotal for a full appreciation of the eye fixation duration results for two-character words. First, information uptake is not limited to the character that is fixated on. When a fixation is on the first character, parafoveal preview allows for simultaneous processing of the second character, and vice versa. We thus found both forward and backward parafoveal preview effects. Despite the fact that initial fixations on the first character typically preceded initial fixations on the second character, processing for both characters thus is by no means strictly serial in nature.

Second, from an information-theoretic perspective, frequency is inversely proportional to the amount of information provided by a character. Hence, the pattern of results for PC1 and PC2 is indicative of a dynamic search for the locus of information. Fixations on the first character are shorter when the first character provides less information (i.e., has a high frequency) and when the second character provides more information (i.e., has a low frequency). Similarly, fixations on the second character are shorter when the first character provides more information (i.e., has a low frequency) and when the second character provides less information (i.e., has a high frequency). Resources are thus allocated dynamically, based on the amount of information provided by both characters.

Recall that for one-character words we observed an effect of the visual complexity of the character. For two-character words we similarly observed an effect of the visual complexity of both the first and the second character. The main effect of PC4, which has high loadings for predictors that describe the visual complexity of the first character is significant in all time windows from -400 to 400 ms, as well as for the time window from 800 to 1000 ms. PC3 is the counterpart of PC4 for the second character. The main effect of PC3 is significant in the -200 to 0, 0 to 200, 400 to 600 and 800 to 1000 ms time windows. Therefore, the onset of the main effect of the visual complexity of the first character is somewhat earlier than the onset for the corresponding effect for the second character. This is unsurprising, because the first character was typically fixated on first. As was the case for PC1 and PC2, however, the qualitative nature of the effects of both PC3 and PC4 is best described through the interaction of both predictors with X Position.

Figure 3.56 presents the interaction of PC4 (left panel; $F = 25.165$, $p < 0.001$) and PC3 (right panel; $F = 28.593$, $p < 0.001$) with X Position for fixations that start between 200 and 400 ms after stimulus onset. Both interactions are significant at a large number of additional time windows. We also observed an interaction of PC4 with X Position for the −400 to −200, −200 to 0, 0 to 200, 200 to 400 and 800 to 1000 ms time windows, whereas the interaction of PC3 with X Position was significant in the −200 to 0, 0 to 200, 200 to 400 and 400 to 600 ms time windows as well.

As can be seen in Figure 3.56, the effects for the visual complexity of the first and second character are qualitatively similar to the vertical mirror images of the effects for the frequency of both characters. Again, the results are best understood in terms of the participant's search for information. The greater the visual complexity of a character, the more information it provides. More complex first characters, therefore, lead to longer fixations on the first character and shorter fixations on the second character. Conversely, more complex second characters result in shorter fixation durations on the first character and longer fixation durations on the second character. These findings are in line with previous findings by Miwa et al. (2014), who found a similar pattern of results for fixation durations of fixation on two-character Japanese words in a lexical decision task.

As was the case for the effects of character level frequency measures, the effects of the character level visual complexity measures demonstrate that parafoveal preview allows for lexical properties of the second character to influence processing during

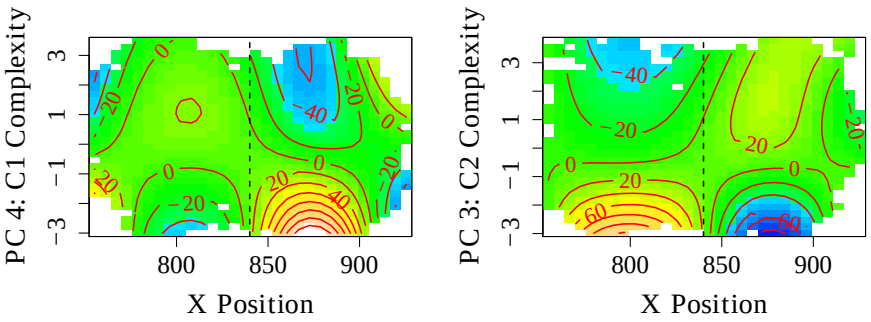

Figure 3.56: Fixation duration results: two-character words. Interaction with horizontal fixation position for character 1 (left panel) and character 2 complexity (right panel) for fixations that start between 200 and 400 ms after stimulus onset. The main effect of horizontal fixation position is omitted from both contour plots for ease of interpretation.

fixations on the first character, and vice versa. Furthermore, the pattern of results for the visual complexity measures provides additional support for the idea that eye fixation patterns are co-determined by a dynamic search for information. Fixation durations are shorter when the character that is fixated on provides less information, as well as when the other character provides more information.

As was the case for the effects of character frequency, the effect sizes of the effects of the visual complexity of the first and second character are strongest for fixations on the other character. For the visual complexity of the first character, the effect size is 69 ms for fixations on the first character, whereas it is 222 ms for fixations on the second character. The effect size for the visual complexity of the second character is 141 ms for fixations on the second character and 174 ms for fixations on the first character. Therefore, fixation durations for the current character seem to be primarily determined by lexical properties of the other character.

In addition to the interaction of visual complexity with `X Position`, we furthermore observed an early interaction between the visual complexity and the frequency of the first character. This interaction between `PC1` and `PC4` was significant in the −200 to 0 and 0 to 200 ms time windows. The effect is depicted for fixations that start in the −200 to 0 ms time window ($F = 10.377$, $p < 0.001$) in Figure 3.57.

For first characters with a high frequency, fixation durations are shorter when the visual complexity of the first character is low (effect size: 60 ms). Therefore, highly frequent simple characters allow for efficient processing of the first characters and a rapid refixation on the second character. For first characters with a low fre-

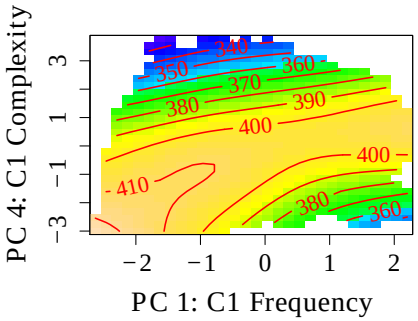

Figure 3.57: Fixation duration results: two-character words. Interaction of character 1 frequency with character 1 complexity for fixations that start between -200 and 0 ms after stimulus onset. Additive contour surface for the main effect of character 1 frequency, the main effect of character 1 complexity and the interaction between character 1 frequency and character 1 complexity.

quency, the effect of visual complexity is reversed, with a greater visual complexity leading to shorter fixation durations (effect size: 88 ms). One possible explanation for the reversal of the visual complexity effect for characters with a low frequency is that the additional visual information provided by characters with a high visual complexity helps process low frequency characters faster. Alternatively, our participant attempted to resolve the increased uncertainty associated with characters with a low frequency and a high complexity by moving onto the second character faster.

In addition to the effects of frequency at the character level, we observed frequency effects at the word level. Word frequency is encoded in `PC5`, which was significant for all time windows between 600 and 1400 ms. The main effect of `PC5` is presented for the 600 to 800 ms time window ($F = 12.006$, $p < 0.001$) in the left panel of Figure 3.58. Consistent with the frequency effect for one-character words, fixation durations are longer for more frequent words. The effect is linear or near-linear and has an effect size of 55 ms.

The main effect of word frequency is first significant for fixations that start between 600 and 800 ms after stimulus onset. Compared to the onset of the character frequency effects, the onset of the main effect of word frequency is late. The relatively late onset of the word frequency effect is unsurprising. Each character in Mandarin Chinese is multiply ambiguous. Only when characters occur together, they contrast the search space enough to be able to zoom in on the intended meaning. Hence, the script enforces a reading strategy that is similar to lexical processing in auditory comprehension.

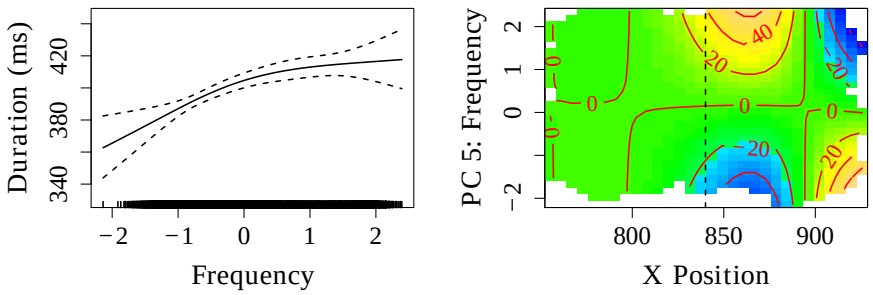

Figure 3.58: Fixation duration results: two-character words. Main effect of whole-word frequency for fixations that start between 600 and 800 ms after stimulus onset and the interaction of whole-word frequency with horizontal fixation position for fixations that start between 200 and 400 ms after stimulus onset. The main effect of horizontal fixation position is omitted from the right panel for ease of interpretation.

We did find some evidence for an earlier onset of the word frequency effect. The right panel of Figure 3.58 shows the interaction between `PC5` and `X Position`, which was significant for the 200 to 400 ms time window only ($F = 12.614$, $p < 0.001$). As noted above, fixation positions are evenly distributed between the first and second character in this time window (see Figure 3.48). At this point in time, the main effect of word frequency, however, is present for fixations on the second character only (effect size: 114 ms).[3] This provides further evidence for the hypothesis that information about the co-occurrence of characters is pivotal for reducing uncertainty about the identity of a two-character word.

Above we described frequency effects at or above the character level. The GAMs for the fixation durations furthermore revealed effects of three frequency measures below the character level. First, we observed main effects of the frequency of the semantic radical of the first (`PC22`) and second character (`PC23`). The main effect of `PC22` reached significance for fixations that start in the −200 to 0 ms time window ($F = 22.026$, $p < 0.001$; effect size: 25 ms) and is depicted in the left panel of Figure 3.59. Fixation durations are longer when the frequency of the semantic radical of the first character is high. The right panel of Figure 3.59 presents the effect of `PC23`, which was significant in the 0 to 200 ms time window ($F = 10.044$, $p < 0.001$; effect size: 28 ms). The effect of `PC23` is in the opposite direction of the effect of `PC22`, with shorter fixation durations when the semantic radical of the second character has a high frequency.

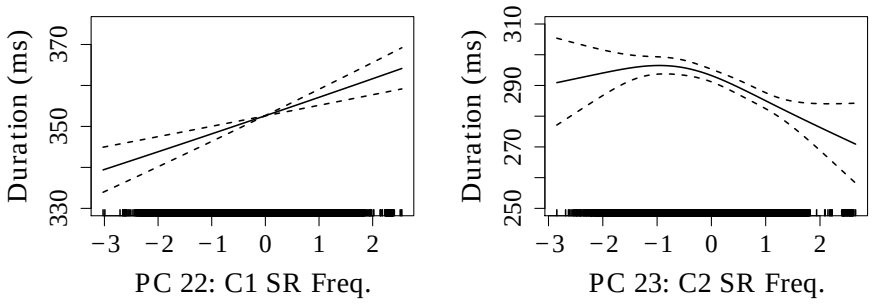

Figure 3.59: Fixation duration results: two-character words. Effects of character 1 SR frequency for fixations that start between -200 and 0 ms after stimulus onset (left panel) and character 2 SR frequency for fixations that start between 0 and 200 ms after stimulus onset (right panel).

[3]Note that we again see a reversal of the effect near the right edge of the analysis window. Given the unreliability of GAMs near the edges when k is set to 4, however, this reversal is not statistically robust.

A large majority of the fixations in both the −200 to 0 (83.23%) and the 0 to 200 ms (98.12%) time window are on the first character. Above, we saw that for fixations on the first character, fixation durations decreased as a function of the first character frequency and increased as a function of the second character frequency. We therefore found an opposite pattern of results for the frequency of the character and the frequency of the semantic radical on fixation durations. The pattern of results observed here thus is consistent with the opposite results for the frequency of the character and the semantic radical in the GAM analysis for the naming latencies.

The main effects of `PC22` and `PC23` were present for fixations that start near stimulus onset. In addition to these main effects, we found later interactions between `PC22` and `X Position` in the 400 to 600 ms time window and between `PC23` and `X Position` in the 200 to 400 ms time window. The interaction between `PC22` and `X Position` showed an effect of `PC22` for fixations on the left character only, with *shorter* fixation durations for high values of `PC22`. In the 400 to 600 time window, however, a mere 8.91% of all fixations are on the left character. The data thus were sparse in the area where this effect was observed. Furthermore, the effect was most prominent near the left edge of the analysis window. The interaction between `PC23` and `X Position` similarly showed an effect near the (right) edge of the analysis window only. Therefore, the statistical reliability of the interactions of `PC22` and `PC23` with `X Position` is questionable. For this reason, plots of the interactions of `PC22` and `PC23` with `X Position` are not shown.

The third frequency effect below the character level is an effect of `PC8`, which encodes the frequency of the diphones for the second character. The effect of `PC8` is significant in the 400 to 600 ms time window ($F = 11.002$, $p < 0.001$) and is shown in Figure 3.60. Fixation durations are shorter when the diphone frequencies for the second character are higher. The effect is present for medium-to-high predictor values only and has an effect size of 56 ms.

In the 400 to 600 ms time window, 91.09% of all fixations are on the second character. Therefore, the main effect of `PC8` in this time window is almost exclusively an effect of `PC8` on fixation durations for fixations on the second character. Above, we observed that fixations durations for fixations on the second character decreased as a function of the frequency of the second character. Hence, the effect of diphone frequency of the second character is in line with the effects of character level frequency measures.

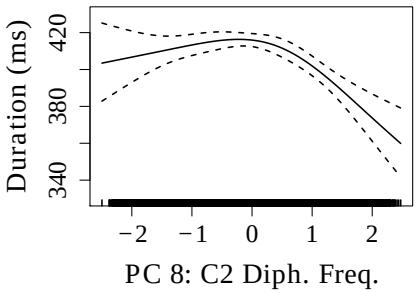

Figure 3.60: Fixation duration results: two-character words. Effect of character 2 diphone frequency for fixations that start between 400 and 600 ms after stimulus onset.

Thus far, we discussed the effects of frequency and visual complexity. The GAM analysis furthermore revealed two effects related to the consistency of the mapping between orthography and phonology. The first effect is an effect of the number (and frequency) of homophones of the first character, which is encoded in `PC10`. The lexical predictors with the highest loadings on `PC10` are `Character 1 Homophones (Tokens)` (0.956), `Character 1 Homophones Frequency` (0.939) and `Character 1 Homophones (Types)` (0.866). The effect `PC10` was significant for the 0 to 200 ms time window ($F = 21.120$, $p < 0.001$) and is depicted in Figure 3.61. The effect is linear, with predicted fixation durations being 38 ms longer for the highest predictor values as compared to the lowest predictor values.

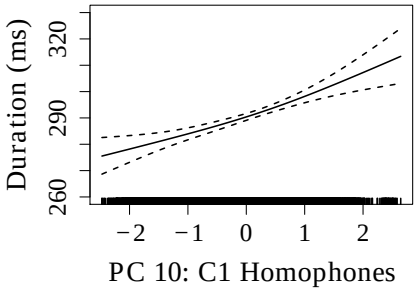

Figure 3.61: Fixation duration results: two-character words. Character 1 homophones for fixations that start between 0 and 200 ms after stimulus onset (right panel).

Almost all fixations (98.12%) in the 0 to 200 time window are on the first character. Therefore, the main effect of PC10 is primarily an effect for fixations on the first character. The results for the eye fixation durations on two-character words we have seen so far suggest that longer fixation durations on the first character for lexical properties of the first character correspond to greater processing costs. The effect of the number of homophones of the first character, hence, is as expected: the more homophones the first character has, the greater the expected processing costs.

The second orthography-to-phonology consistency measure that showed a significant effect is PC13, which describes the number of homographs for the second character and their frequency (highest loadings: Character 2 Homographs (Tokens) (0.964), Character 2 Homographs (Types) (0.959) and Character 2 Homographs Frequency (0.938)). No main effect of PC13 was observed. In the 400 to 600 ms time window, we observed a significant interaction of PC13 with X Position ($F = 5.696$, $p < 0.001$). However, an effect was present at the left edge of the analysis window only. In the 400 to 600 ms time window a mere 8.91% of all fixations were on the left character. Therefore, the effect of PC13 is likely to be an outlier effect that is not statistically robust. For this reason, a plot of the effect of PC13 is not shown.

The GBM fitted to the fixation durations showed high relative influences for the entropy of both the first and the second character. The importance of the entropy of both characters is reflected in the GAM analysis, which shows significant effects of both PC42 (character 1 entropy) and PC39 (character 2 entropy) in the 0 to 200 ms time window. The effect of PC42 ($F = 39.830$, $p < 0.001$) is presented in the left panel of Figure 3.62. The effect is linear, with shorter fixation durations when the entropy of the first character is high (effect size: 49 ms). By contrast, when the entropy of the second character is high, fixation durations are longer ($F = 10.325$, $p < 0.001$; effect size: 78 ms).

The effects of PC39 and PC42 fit well with the results reported above. To appreciate this, we need to combine three observations made thus far. First, the GAM analysis on the naming latencies revealed that naming latencies were shorter when character 1 or character 2 had a high entropy. Second, lexical characteristics of a character that lead to longer naming latencies result in longer fixations on that character and shorter fixations on the other character. Conversely, lexical characteristics of a character that lead to shorter naming latencies result in shorter fixations on that character and longer fixations on the other character. Third, nearly all fixations

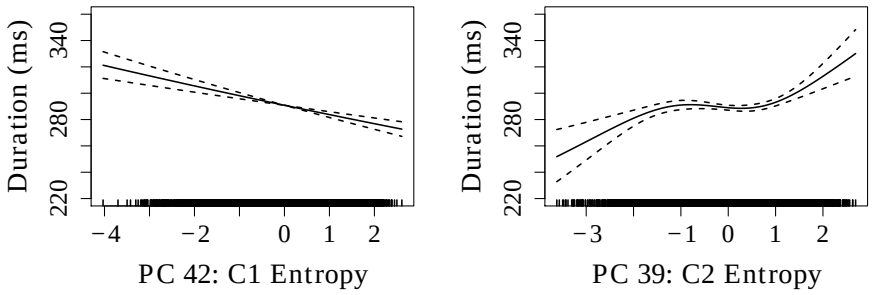

Figure 3.62: Fixation duration results: two-character words. Effects of character 1 entropy (left panel) and character 2 entropy (right panel) for fixations that start between 0 and 200 ms after stimulus onset.

(98.12%) that start in the 0 to 200 ms time window are on the left character.

Now consider the effect of `PC42` in the left panel of Figure 3.62. Naming latencies were shorter for high values of `PC42` (i.e., high character 1 entropy). Given the fact that most fixations for the depicted 0 to 200 ms time window are on the first character, we expect a similar effect of `PC42` on the fixation durations. This is exactly what we find.

For `PC39` we likewise observed shorter naming latencies for high predictor values. In contrast to `PC42`, however, `PC39` describes a lexical property of the second character. For fixations on the first character, we thus expect an opposite effect of `PC39` on the naming latencies. As can be seen in the right panel of Figure 3.62, this prediction is borne out: fixation durations in the 0 to 200 ms time window are longer when the second character has a high entropy. The effect of the entropy of the second character during fixations on the first character is a further demonstration of the fact that parafoveal preview allows for joint processing of the first and second character.

In addition to the main effect of `PC42` in the 0 to 200 ms time window, we observed an interaction of `PC42` with `X Position` in the 200 to 400 ms time window ($F = 11.681$, $p < 0.001$). As can be seen in Figure 3.63, the main effect of `PC42` is present for fixations on the left character only (effect size: 62 ms).

The GAM analysis furthermore revealed an effect of `PC32`. The lexical predictor with the highest loading on `PC32` is `Character 2 RE`, with a loading of 0.977. The absolute values for all other loadings on `PC32` are smaller than 0.20. `PC32` therefore describes how similar the frequency distribution of first characters for a given second character is to the distribution of first characters across all two-character words in

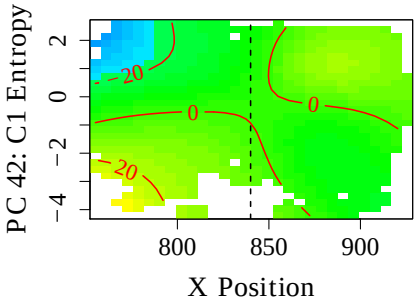

Figure 3.63: Fixation duration results: two-character words. Interaction of horizontal fixation position with character 1 entropy for fixations that start between 200 and 400 ms after stimulus onset. The main effect of horizontal fixation position is omitted for ease of interpretation.

the CLD. The effect of `PC32` was significant for fixations that start between 200 and 400 ms after stimulus onset ($F = 21.116$, $p < 0.001$). As can be seen in Figure 3.64, the effect of `PC32` is linear, with shorter fixation durations when the distribution of character 1 frequencies for the second character is less similar to the distribution of character 1 frequencies across all two-character words.

Unlike most effects we have seen thus far, the effect of `PC32` is independent of the horizontal fixation position. Fixations in the 200 to 400 ms time window are evenly split between the first (47.97%) and the second (52.03%) character. Furthermore, a post-hoc verification in which we added the interaction between `PC32` and `X Position` to the model confirmed that there is no reason to believe that the effect of `PC32` is different for fixations on the left character and fixations on the right character ($F = 1.607$, $p = 0.224$). Fixation durations thus are shorter when the second character

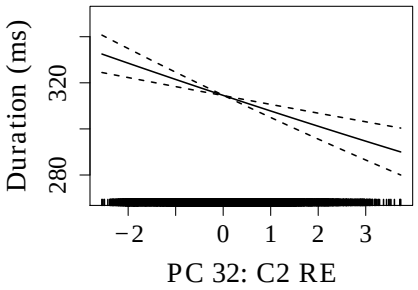

Figure 3.64: Fixation duration results: two-character words. Effect of character 2 relative entropy for fixations that start between 200 and 400 ms after stimulus onset.

combines with first characters in an unusual manner, independent of the horizontal position of a fixation.

The final numerical predictor that showed a significant effect on fixation durations is `PC36`, which encodes the entropy over the left and right character frequency. The main effect of `PC36` was significant in the −200 to 0 ms time window ($F = 8.102$, $p < 0.001$, effect size: 50 ms). As can be seen in the left panel of Figure 3.65, fixation durations are longer when the entropy over the character frequencies is higher.

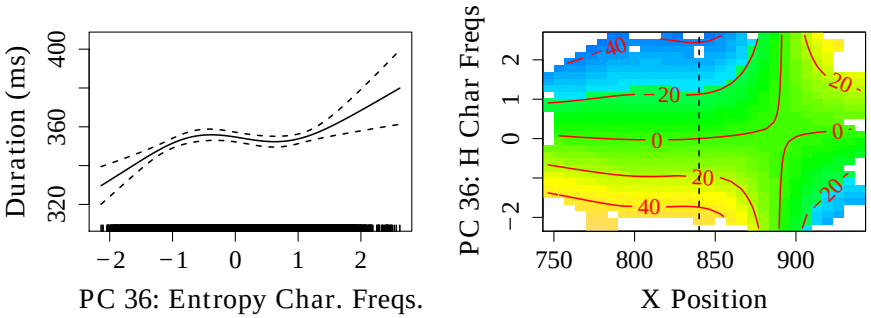

Figure 3.65: Fixation duration results: two-character words. Main effect of entropy character frequencies for fixations that start between -200 and 0 ms after stimulus onset (left panel) and the interaction of horizontal fixation position with entropy character frequencies for fixations that start between 400 and 600 ms after stimulus onset (right panel). The main effect of horizontal fixation position is omitted from the right panel for ease of interpretation.

However, in the 400 to 600 ms time window we observed an interaction between `PC36` and `X Position` ($F = 8.797$, $p < 0.001$). This interaction is presented in the right panel of Figure 3.65. For all but the most rightward horizontal fixation positions, the effect of `PC36` is reversed as compared to the main effect of `PC36` in the −200 to 0 time window. For the most rightward fixation positions the effect is qualitatively similar to the main effect of `PC36` in the −200 to 0 ms time. Again, the robustness of the interaction between `PC36` and `X Position` is questionable. An overwhelming majority of the data points falls in the green region in the right half of the plot. Therefore, for most fixations there is no evidence for an effect of `PC36` in either direction in the 400 to 600 ms time window.

While the GBM analysis suggested that `Character 2 Structure` co-determines fixation durations to a considerable degree, the GAM analysis provided no evidence for an effect of `Character 2 Structure`. Therefore, the effect of `Character 2 Structure` may be limited to restricted areas of the distributional space of Man-

darin Chinese for which other lexical predictors have specific values (i.e, the high relative influence of `Character 2 Structure` in the GBMs may reflect interactions of `Character 2 Structure` with multiple other lexical predictors). The GAM analysis did, however, show effects of three other categorical variables on the fixation durations: `Character 1 Type`, `Character 2 Type` and `Character 2 Tone`.

The effect of `Character 1 Type` was significant in the 0 to 200 ms time window. Fixation durations are shortest for pictologic characters (−34 ms), followed by pictographic (−19 ms), pictosynthetic (−2 ms), other (baseline) and pictophonetic (+0 ms) characters. Pairwise comparisons showed that the pairwise differences between pictologic characters on the one hand and pictophonetic, pictosynthetic and "other" characters on the other hand were significant. Furthermore, the pairwise difference between pictographic and pictophonetic characters reached significance.

`Character 2 Type` showed a significant effect in the 0 to 200 and 200 to 400 ms time windows. For `Character 2 Type`, the order of the character types with respect to fixation duration is reversed. The shortest fixation durations were observed for pictophonetic characters (+4 ms in the 0 to 200 ms time window; −12 ms in the 200 to 400 ms time window), followed by other (baseline), pictosynthetic (+13 ms; −4 ms), pictographic (+27 ms; +17 ms) and pictologic (+27 ms; +30 ms) characters. Pairwise differences were significant between pictographic characters on the one hand and pictophonetic and "other" characters on the other hand for the 0 to 200 ms time window. For the 200 to 400 ms time window the only pairwise difference that reached significance was the pairwise difference between pictophonetic characters and pictographic characters.

Again, we observed an opposite pattern of results for the character 1 and character 2 counterparts of a lexical predictor on early fixation durations. Given that early fixations are primarily on the first character, we conclude that the order of character types for the first character in terms of predicted fixation duration is proportional to processing costs, as is the reverse order of character types for the second character. Pictophonetic characters are therefore associated with the greatest processing costs, whereas pictologic and pictographic characters are processed with relative ease. The results of character type are thus in line with the effect of `Character 2 Type` on the naming latencies, which showed longer naming latencies for pictophonetic characters as compared to pictographic characters.

Finally, we found an effect of `Character 2 Tone` in the 400 to 600 ms time window. Unsurprisingly, this is the time window in which a large majority of pronunciations begin (as noted in our discussion of the effect of `Final Phoneme` on fixation durations, 87.08% of all naming latencies are between 400 and 600 ms). Pairwise differences were significant between Tone 3 (−30 ms) on the one hand and Tone 1 (reference level) and Tone 2 (−7 ms) on the other hand.

For pronunciation durations, we observed an effect of `Character 2 Tone` as well, with pronunciation durations being the shortest for Tone 3 and the longest for Tone 1 and Tone 2. The order of the tones with respect to pronunciation duration thus is remarkably similar to the order of the tones with respect to fixation duration. This suggests that the effect of tone on the pronunciation durations is at least partly due to lexical differences between words and characters with different tones, rather than to differences in the acoustic realizations of tones only.

3.5 Generalizability: a second participant

The current study is a single-participant study. For Vietnamese, Pham and Baayen (2015) (see also Pham, 2014) showed that the results of a large-scale single-subject lexical decision study were qualitatively similar to the results of a multiple-subject lexical decision experiment carried out on a smaller scale. However, the ability of the single-subject experiment to uncover relatively subtle effects was greater than that of the multiple-subject study. Nonetheless, while the analyses described above provide a detailed overview of lexical processing of the participant under investigation, the experimental design raises some questions regarding the extent to which the results reported here generalize to other speakers of Mandarin Chinese.

To gain further insight into the generalizability of the results of our single-subject study reported above to other (highly educated) speakers of Mandarin Chinese we gathered naming latencies for all 30,645 words in the CLD for the author of this dissertation. We analyzed this second set of naming latencies in the same manner as described above, with GBMs and GAMs fitted to the naming latencies and pronunciation durations for each word. No eye fixations were recorded for the author of this dissertation. Therefore, an analysis of the eye fixation durations was not included for this participant.

The results of the analysis on the naming latencies were comparable with those observed for our main participant. The naming latency models were dominated by strong frequency effects, both at the character level and at the word level. The observed frequency effects were qualitatively similar to the frequency effects reported above, with shorter naming latencies for more frequent words and characters. Furthermore, we found robust effects of character 1 and character 2 entropy that were qualitatively similar to the entropy effects reported above. Furthermore, the effects of phonological measures on the naming latencies were more subtle, although we did observe an effect of phonological neighbourhood density that was qualitatively similar to the effect of phonological neighbourhood density reported above.

In addition to these similarities, we also observed some differences between the two participants. The author of this dissertation was born in Taiwan, which means that she is a native reader of traditional Chinese, rather than simplified Chinese. This influenced her pattern of results for the naming latencies in two ways. First, the effects of traditional frequency were much stronger for the author of this dissertation than for the participant described above. This is unsurprising and highlights the importance of using a corpus that is representative for the reading experience of a participant.

Second, the effects of visual complexity were much weaker for the author of this dissertation as compared to the participant described above. For our main participant, visual complexity showed the expected effect, with longer naming latencies for words that consist of visually complex characters. For native readers of traditional Chinese, however, a second factor enters the equation when the effects of visual complexity are considered: familiarity with the visual input. Characters that have gone through the process of simplification are visually less complex, but unfamiliar to native readers of traditional Chinese. By contrast, characters that have not been simplified are visually complex, but familiar to native readers of traditional Chinese. Hence, familiarity with the visual input is positively correlated with visual complexity. For the data for the author of this dissertation reported here, this resulted in decreased effect sizes of visual complexity, with a statistically significant effect of visual complexity for the first character in two-character words only.

The above dissimilarities were straightforward consequences of the differences between simplified and traditional Chinese. We furthermore found two differences between the pattern of results for the two participants that are relatively independent of the simplified versus traditional Chinese issue. First, for our main participant

we observed relatively few effects of categorical variables, with an effect of `Character 2 Type` for two-character words only. For the author of this dissertation, we observed additional effects of `Character 1 Type` and `Character 1 Tone` for both one-character words and two-character words. Second, whereas we found significant effects of the frequency and complexity of the semantic radical (of the first character) for our main participant, the analysis of the data for the author of this dissertation revealed no effects of lexical properties of the semantic radical. Therefore, whether or not the semantic radical plays a role in lexical processing may vary between participants.

The pronunciation duration analyses for the author of this dissertation showed a large number of effects of phonological measures, as was the case for the pronunciation duration analyses for our main participant reported above. For both participants, we observed effects of the number of phonemes in both characters, with longer pronunciation durations for words that consist of characters with a larger number of phonemes. Furthermore, the analyses revealed a number of phonological frequency effects. As before, we observed effects of diphone frequency for both the first and the second character. These effects were qualitatively similar to the diphone frequency effects observed above. For our main participant we observed effects of the phoneme frequency of the first (and only) character of one-character words and of the second character of two-character words. We did not find statistically significant evidence for such effects for the author of this dissertation. We did, however, observe an effect of the phoneme frequency of the first character of two-character words that was qualitatively similar to the effect of phoneme frequency for one-character words for our main participant.

For our main participant we also found effects of the phonological neighbourhood of both the first and the second character. For the author of this dissertation, the effect of phonological neighbourhood was limited to the second character. Interestingly, this effect of phonological neighbourhood density was in the opposite direction of the effect of phonological neighbourhood density of the second character reported above, with shorter naming latencies for second characters that live in dense phonological neighbourhoods. We conclude that while phonological neighbourhood characteristics seem to influence pronunciation durations in word naming, individual differences may exist with respect to the nature of phonological neighbourhood effects. Therefore, additional research is necessary to gain further insight into the nature of these effects across different (groups of) participants.

The analyses for our main participant had shown that frequency effects on pronunciation durations were most prominent at the word level. The analysis for the author of this dissertation revealed a similar pattern of results. The GAM analyses revealed word frequency effects for one and two-character words that were qualitatively similar to the effects of word frequency reported above, with shorter pronunciation durations for more frequent words.

Finally, we observed effects of the tone of both the first and the second character for both participants. The effect of the tone of the first character was most prominent for one-character words. For both participants, we observed relatively long pronunciation durations for tones 2 and 3 and shorter pronunciation durations for tones 1 and 4. The difference between the predicted values for the fastest and the slowest tone was larger for the author of this dissertation (131 ms) than for our main participant (97 ms). For the second character in two-character words, pronunciation durations were longest for tone 1 and 2 and shortest for tones 3 and 4 for both participants. Again, the magnitude of the effect was larger for the author of this dissertation (134 ms) as compared to our main participant (38 ms).

The increased influence of tone on pronunciation durations for the author of this dissertation likely results from a difference in pronunciation strategy for both participants. The author of this dissertation pronounced words in a formal manner, paying close attention to acoustic detail. By contrast, our main participant focused on responding as fast as possible and paid less attention to acoustic detail. As a result, naming accuracy was somewhat higher for the author of this dissertation (96.04%) as compared to our main participant (94.37%). Furthermore, pronunciation durations were longer for the author of this dissertation (mean pronunciation duration one-character words: 373; mean pronunciation duration two-character words: 567 ms) as compared to our main participant (mean pronunciation duration one-character words: 316 ms; mean pronunciation duration two-character words: 472). The longer pronunciation durations for the author of this dissertation resulted in stronger manifestations of the inherent differences in acoustic durations for different tones.

In conclusion, the analyses of the naming latencies and pronunciation durations for both participants revealed similar patterns of results. The naming latency analyses for both participants were dominated by frequency effects, but showed robust effects of entropy as well. Pronunciation durations for both participants were primarily influenced by phonological properties of a word and its characters. Hence, we expect the most prominent effects reported here to generalize to other highly educated speakers of simplified Chinese.

We observed a number of differences between the two participants as well. Some of the most striking differences are likely to reflect systematic differences between lexical processing in simplified Chinese for native readers of simplified Chinese and native readers of traditional Chinese and highlight the importance of using native readers of simplified Chinese when investigating lexical processing in simplified Chinese. Other discrepancies between the analyses for both participants may be individual differences that are independent of the linguistic background of a participant. As far as some of the more subtle effects reported here are concerned, therefore, it is important to verify that these effects are present for other readers of simplified Chinese as well in future research.

3.6 General discussion

In this chapter we investigated the explanatory power of the lexical predictors in the CLD in the word naming task through a single-subject study for all 30,645 words in the CLD. In this single-subject study we investigated the effects of the lexical predictors in the CLD on three behavioural measures of language processing collected during the word naming task: naming latencies, pronunciation durations and eye fixation durations. The data were analyzed using gradient boosting machines (GBMs) and generalized additive models (GAMs) to obtain a detailed picture of the quantitative and qualitative effects of the predictors on lexical processing in the word naming task.

Frequency measures co-determined all three dependent variables. The naming latency analyses were dominated by frequency-related measures, both at the word level and at the character level. The effects of frequency in these analyses were as expected, with shorter naming latencies for words and characters with a higher frequency (c.f., Y. Liu et al., 2007; Y. N. Chang et al., 2016; Lee et al., 2015; Sze et al., 2014; Zhang & Peng, 1992; Peng et al., 1999). Frequency measures also co-determined eye fixation durations over time (G. Yan et al., 2006). For both first and second characters, the effect of character frequency was first significant for fixations that started as early as 200 to 0 ms *before* stimulus onset.

Parafoveal preview allowed for bidirectional joint processing of both characters. We found effects of the frequency of the second character for fixations on the first character, and vice versa. Fixations on a character were less long for frequent characters than for infrequent characters. By contrast, fixations on a character were longer when the other character was frequent than when the other character was

infrequent. Frequent words provide less information than infrequent words. The eye fixation patterns on two-character words thus are co-determined by a dynamic search for information. The greater the amount of information provided by a character, the longer it is fixated on and the more quickly attention is directed towards that character.

The effect of word frequency on eye fixation durations was first significant for fixations that started between 200 and 400 ms after stimulus onset and was most prominent for fixations that started even later. We thus found a prolonged process of uncertainty reduction with respect to the identity of a two-character word, which is a straightforward consequence of the fact that each character is multiply ambiguous in Mandarin Chinese.

Pronunciation durations were influenced by frequency measures as well. The GAM analysis revealed a significant effect of word frequency, with shorter fixation durations for high frequency words as compared to low frequency words. The GBM analysis furthermore provided some evidence for an effect of the frequency of the second character. The GAM analysis, however, revealed no such effects.

In addition to frequency effects at or above the character level, we found some evidence for frequency effects below the character level. The effect of semantic radical frequency on naming latencies was mild and limited to the first character. The effect of semantic radical frequency on eye fixation durations, by contrast, was present for both the first and the second character, although the effect sizes of these effects were limited as well. Furthermore, the analyses of the naming latencies for a second participant did not reveal effects of semantic radical frequency.

The effect of the semantic radical frequency of the first character on the eye fixation durations is first significant for fixations that start between 200 and 0 ms before stimulus onset. The temporal onset of the effect of the semantic radical frequency of the first character thus coincides with the onset of the first character frequency effect. The onset of the effect of the semantic radical frequency of the second character (0 to 200 ms after stimulus onset) is somewhat later than the onset of the effect of the frequency of the second character (200 to 0 ms before stimulus onset).

We conclude that – to the extent that semantic radical frequency effects characterize lexical processing in Mandarin Chinese – these effects do not seem to precede frequency effects at the character level, at least for the second character. The current findings therefore do not straightforwardly support a visual information uptake

process that starts at a small grain size (i.e., semantic radicals or smaller) and then integrates the information at this grain size (c.f., Taft, 2006; Taft et al., 1999; Taft & Zhu, 1997) to arrive at a correct understanding of the character is hence unlikely. Instead, the current pattern of results suggests that optimal processing at the character level is holistic from the start. Specific attention to sub-character level visual components may be necessary if and only if initial holistic processing of the character proved insufficient for character discrimination and not for all readers.

A second benchmark effect in word naming tasks across languages is the effect of visual complexity. As expected, we found robust effects of visual complexity on the naming latencies, with longer naming latencies for more complex characters (c.f., Y. Liu et al., 2007; Y. N. Chang et al., 2016; Leong et al., 1987; Lee et al., 2015). The eye fixation duration analysis similarly showed effects of visual complexity, with early effects of both character 1 complexity (first significant for fixations that started 400 to 200 ms *before* stimulus onset) and character 2 complexity (first significant for fixations that started 200 to 0 ms *before* stimulus onset). Again, eye fixation patterns were characterized by a dynamic search for information. Visually complex characters were fixated on longer than visually simple characters and fixations durations on a character were shorter when the other character had a greater visual complexity. The effects of visual complexity on the pronunciation durations were more subtle. Nonetheless, both the GAM and the GBM analysis revealed an effect of the picture size (in bytes) of the second character.

The principal component regression analysis using GAMs provides little insight into the exact visual features that drive complexity effects. The GBM analyses, by contrast, provide information about the relative contribution of individual predictors to the models. Interestingly, different types of visual complexity measures were most predictive for different dependent variables. As noted above, picture size co-determined pronunciation durations. Naming latencies, by contrast, were influenced primarily by stroke counts. Finally, the strongest visual complexity measures in the GBM fitted to the eye fixation durations were pixel counts. Different properties of the visual input therefore co-determine lexical processing at different points in time.

Whereas naming latencies were primarily influenced by frequency and – to a lesser extent – visual complexity, the models for pronunciation durations were dominated by phonological properties of the words and characters. Three types of phonological measures proved particularly predictive. First, we found a large number of phonological frequency effects, both at the phoneme level and at the diphone level.

Typically, pronunciation durations were longer for characters with more frequent phonemes and diphones. We argued that this is a straightforward consequence of learning theory, which predicts that associations between characters and phonological units are inversely proportional to the number of words in which these units appear (see e.g., Baayen et al., 2011).

Second, we found effects of phonological complexity. As expected, naming latencies were longer for characters with a greater number of phonemes. Third, the phonological neighbourhood density of both characters influenced pronunciation durations, both in the GBM analysis and in the GAM analysis. The effect of phonological neighbourhood density for a second participant, however, was in the opposite direction of the corresponding neighbourhood density effect for our main participant. Further research is required to establish the exact nature of phonological neighbourhood effects, which may vary between participants.

The measures discussed thus far describe lexical properties of the word or character in isolation. The various analyses for two-character words, however, revealed that the combinatorial properties of characters are essential for efficient processing. We found robust effects of the entropy of the first and second character on both naming latencies and eye fixation durations, with shorter naming latencies and shorter fixation durations for high entropy characters. The entropy effects observed here thus are in the opposite direction of the entropy effects typically observed for English. One potential explanation for the facilitatory effects of entropy observed here is that the orthography-to-phonology mapping is more consistent for characters that combine with many other characters as compared to characters that combine with few other characters.

To our knowledge, the entropy effects reported here are the first entropy effects reported for simplified Chinese. The entropy effects on eye fixation durations are first significant for fixations that start between 0 and 200 ms after stimulus onset. The combinatorial properties of characters thus influence lexical processing after the characters are accessed, but before the word is discriminated. The effects of character 1 and character 2 entropy are complemented by facilitatory effects of the trigram entropy and relative entropy of the first character and an inhibitory effect of the entropy over the character frequencies on the naming latencies.

Above we mentioned that we found some evidence for the involvement of semantic radicals, with moderate effects the frequency of semantic radicals on naming latencies and eye fixation durations. We similarly found some evidence for the involvement of phonetic radicals, particularly in the naming latencies for one-character

words. For both participants we observed a strong facilitatory effect of the number of characters in which the same phonetic radical is pronounced the same (c.f., Y. Liu et al., 2007; Seidenberg, 1985; Hue, 1992). Particularly in the absence of disambiguating information provided by a second character, therefore, the phonetic radical helps reduce uncertainty about the correct pronunciation of a character.

Nevertheless, the overall influence of orthography-to-phonology consistency measures was modest given the fact that homographs and homophones are ubiquitous in simplified Chinese. The abundance of entropy effects suggests that the uncertainty introduced by shared orthographic and phonological forms is at least partially resolved at a different level of lexical processing: the combination of characters into words. Once the combination of characters in the current word is established, most of the uncertainty about the orthography-to-phonology mapping is resolved: above the character level much fewer orthographic or phonological forms are shared.

Although the research presented here provides a detailed overview of the lexical properties that influence language processing in the word naming task, the current study has some shortcomings. First, we exclusively focused our attention on the processing of one-character words and two-character words presented in isolation. Natural language processing, however, is rarely limited to single words. Instead, we typically encounter words in the context of other words. Given the prominence of entropy effects for the current data, the question arises if and to what extent entropy effects manifest themselves above the word level as well. Furthermore, the current data showed that eye fixation patterns are characterized by joint processing of multiple linguistic elements through bidirectional parafoveal preview and by a dynamic search for information. Again, the question arises if and to what extent this result generalizes to lexical processing above the word level.

In addition, the current study was limited to a single linguistic task: word naming. The word naming task has a long and fruitful tradition in psycholinguistics. It therefore provides an excellent starting point for the investigation of lexical processing in Mandarin Chinese, as well as for the evaluation of the measures in the CLD. The knowledge obtained here, however, needs to be complemented by data gathered through other linguistic tasks to obtain a comprehensive understanding of lexical processing in Chinese.

We return to these issues in the next chapter. Despite these shortcomings, however, the results presented here provide a rich and detailed overview of lexical processing in the word naming task for an enormous set of one-character words and

two-character words for a native reader of simplified Chinese. We hope to be able to extend the data presented here to a larger set of participants in future research to establish the generalizability of the results reported here. However, consistent with the findings of Pham and Baayen (2015) for Vietnamese (see also Pham, 2014), the analysis of a second set of naming latencies for a native reader of traditional Chinese suggested that the key effects reported here are likely to generalize to the larger population of highly educated speakers of Mandarin Chinese. The data presented here thus provide valuable information about lexical processing in simplified Chinese.

4
Phrase reading

4.1 Introduction

In Chapter 3, we investigated lexical processing in Mandarin Chinese at the word, character and sub-character levels. In a large-scale single-participant word naming study, we observed effects of a wide range of lexical predictors on naming latencies, pronunciation durations and eye fixation patterns, including effects of word frequency, character frequency, semantic radical frequency, phoneme and diphone frequency, visual complexity, phonological neighbourhood density, phonetic radical consistency and various information-theoretic measures.

Real-life linguistic input, however, rarely consists of isolated words. Instead, words are embedded in phrases, sentences, paragraphs and texts. Successful language processing requires not only lexical access to individual words, but also the integration of words into larger linguistic contexts. The current chapter is an attempt to investigate language processing both at and above the word level through a phrase and sentence reading task.

Below, we report the results of two experiments in which we monitored eye movement patterns during phrase and sentence reading. The phrases and sentences in these experiments revolve around a particular construction in Mandarin Chinese: the locative phrase. Locative phrases are the semantic equivalent of prepositional phrases in English and consist of the semantically empty preposition 在, a GROUND noun and a locative marker that is semantically similar to English prepositions. An example of a locative phrase, for instance, is 在书桌下 ("under the desk"), which consists of the preposition 在, the GROUND noun 书桌 ("desk") and the locative

marker (“下”). Locative markers will henceforth be referred to as topological nominals.

In the word naming study reported in Chapter 3, the robust effects of the information-theoretic measures entropy, trigram entropy and relative entropy were particularly striking. The effects of entropy measures in word naming indicate that the combinatorial properties of characters influence lexical processing in Chinese. Hence, the question arises if and to what extent the way words combine into larger linguistic units influences lexical processing as well.

Locative phrases in Mandarin Chinese have a well-defined form and therefore provide an excellent opportunity to gauge the influence of combinatorial properties above the word level in a consistent linguistic environment. The influence of such higher-level combinatorial properties can be investigated through lexical predictors such as phrase frequency or phrase-level entropy. A particularly interesting lexical predictor for locative phrases, too, is prepositional relative entropy: a measure of how prototypical a noun's use of prepositions is. Previously prepositional relative entropy has been shown to influence lexical processing in English. Both Baayen et al. (2011) and Hendrix et al. (2016) found increased processing difficulties for nouns that use prepositions in an a-typical way. Hendrix et al. (2016) furthermore observed phrase frequency effects for prepositional phrases.

For Mandarin Chinese, prepositional relative entropy is a measure of how prototypical a GROUND noun's use of topological nominals is. Given estimated probabilities p (relative frequencies) of locative expressions for a given GROUND noun and estimated probabilities q of the topological nominals across all GROUND nouns, prepositional relative entropy is defined as:

$$\text{Relative Entropy} = \sum_{i=1}^{n}(p_i * \log_2 (p_i/q_i)) \tag{4.1}$$

where n is the number of GROUND nouns in the language (or, from a more practical perspective, the number of GROUND nouns under investigation).

A fictive example for prepositional relative entropy is presented in Table 4.1. The lexicon for this fictive example contains 6 topological nominals. The relative entropy for the example GROUND noun 书桌 (“desk”) can be calculated through a comparison of the two probability distributions: the probability distribution of topological nominals for locative phrases containing the GROUND noun 书桌 (p) and the probability distribution of topological nominals across all GROUND nouns (q).

Table 4.1: Prepositional relative entropy: fictive example

	书桌 ("desk")		all GROUNDs	
topological nominal	freq.	prob. p	freq.	prob. q
前 ("in front of")	71	0.56	2,223	0.07
上 ("on")	36	0.28	18,865	0.61
旁 ("next to")	14	0.11	689	0.02
下 ("under")	2	0.02	649	0.02
里 ("in")	2	0.02	5,208	0.17
边 ("at the edge of")	1	0.01	3,304	0.11

Entering these probability distributions into Equation 4.1 for our fictive example gives a relative entropy of 1.54 for the GROUND noun 书桌.

The more similar the distributions p and q are, the lower the prepositional relative entropy. The two most frequent topological nominals are "in" (里) and "on" (上), which together comprise 60% of topological nominal tokens in the SCCoW for the GROUND nouns used in this study. GROUND nouns that frequently use "in" (里) and "on" (上), such as clothes (衣服, `RE (SCCoW)`: 0.38) or airplane (飞机, `RE (SCCoW)`: 0.60) have low prepositional relative entropy. High relative entropy GROUND nouns, by contrast, are grounds that frequently use other topological nominals, such as tree (树, `RE (SCCoW)`: 1.74) or woman (女人, `RE (SCCoW)`: 3.54). Previous studies have shown that higher relative entropy leads to greater processing costs (Milin, Filipović Durdević, & Moscoso del Prado Martín, 2009; Milin, Kuperman, et al., 2009; Kuperman et al., 2010; Baayen et al., 2011; Hendrix et al., 2016).

Below, we report the results of two experiments in which participants read locative phrases. In the first experiment locative phrases were presented in isolation. The presentation of locative phrases in isolation allows for a first exploration of locative phrase processing in a highly controlled linguistic environment. The task of reading locative phrases in isolation, however, has limited ecological validity. We therefore carried out a second experiment in which we embedded the locative phrases of the first experiment in two types of sentential contexts: one in which the locative phrase appeared early in the sentence and one in which it appeared late in the sentence. Following our description of the results for both experiments, we discuss the implications of the findings of the current study for our understanding of language processing at the phrase level in Mandarin Chinese.

4.2 Experiment 1

4.2.1 Methods

4.2.1.1 Participants

Twenty-five participants took part in the experiment. All participants were native readers of Mandarin Chinese born in mainland China and living in Tübingen. Their mean age was 25.48 (sd: 3.48). Seventeen participants were female, eight were male. All participants had normal or corrected to normal vision. Participants received 20 euro for their participation.

4.2.1.2 Materials

Ninety-six nouns from Mandarin Chinese were selected to serve as GROUND nouns in the experiment. For each of the GROUND nouns, we constructed four locative phrases, consisting of the semantically empty preposition 在, a GROUND noun and a locative marker (i.e., topological nominal) that is semantically similar to English prepositions. The total number of stimuli therefore was 384.

The experiment was designed and carried out prior to the construction of the CLD. Rather than selecting phrases based on their frequency in the SCCoW or the Gigaword corpus, we therefore selected locative phrases based on their frequency in the Chinese Internet Corpus of Sharoff (2006), which consists of 280 million word tokens. We extracted phrase frequencies for all locative phrases, consisting of the preposition 在, one of the 96 GROUND nouns and a topological nominal from a precompiled list of 51 topological nominals. For a given GROUND noun, the phrases at 25%, 50%, 75% and 100% of the phrase frequency distributions were selected as stimuli for the experiment. This procedure resulted in 40 of the 51 topological nominals being used in one or more experimental items.

The experimental task was a translation verification task. We translated each locative phrase into English. Fifty percent of the translations were correct, the other fifty percent were incorrect. Incorrect translations differed from the locative phrases in Mandarin with respect to either the GROUND noun (e.g., 在柜子里面 ("in the cabinet") translated as "in the bottle") or the locative information (e.g., 在冰箱里面 ("in the fridge") translated as "on the fridge").

4.2.1.3 Design

The experiment consisted of 384 trials, the order of which was randomized between participants. The experimental items were preceded by 5 practice items. We collected three dependent variables related to the eye movement patterns during the experiment: `Fixation Duration`, `Fixation Position` and `Fixation Probability`. `Fixation Duration` is the duration in milliseconds of a fixation. `Fixation Position` is the position of a fixation, measured in the number of pixels from the left edge of the word. `Fixation Probability` is the probability of a fixation on a word.

We included a number of variables as control variables in our design. `Trial` is the trial number within the experiment. `X Position` is the horizontal fixation position, measured in pixels from the left edge of the word. `Y Position` is the vertical fixation position, measured in pixels from the top of the screen. For those fixations for which a previous fixation was present, we included two further control variables: `Previous Fixation Duration` and `Saccade Length`. `Previous Fixation Duration` is the duration in milliseconds of the previous fixation. `Previous Fixation Duration` was log-transformed to remove a rightward skew from the `Previous Fixation Duration` distribution.

`Saccade Length` is the horizontal distance between the previous fixation and the current fixation, measured in pixels. `Saccade Length` could not be included as a predictor when using `Fixation Position` as the dependent variable, because it provides the information that is encoded in `Fixation Position` (correlation between `Saccade Length` and `Fixation Position`: $r = 0.59$). To be precise, `Saccade Length` is the sum of the fixation position in the word and the distance between the previous fixation and the left border of the word. For the dependent variable `Fixation Position` we therefore substituted `Saccade Length` with `Partial Saccade Length`. `Partial Saccade Length` is defined as the distance between the previous fixation and the left border of the word (again measured in pixels). Hence, it excludes the information that is encoded in `Fixation Position` (correlation between `Partial Saccade Length` and `Fixation Position`: $r = 0.14$).

From the CLD we retrieved all word-level predictors, both for the GROUND noun and for the topological nominal. This led to a total of 24 independent variables. The first 12 predictors are frequency and contextual diversity measures from the SCCoW, the Gigaword corpus and SUBTLEX-CH: `Frequency GROUND Noun (SCCoW)`, `Frequency GROUND Noun (Gigaword)`, `Frequency GROUND Noun (SUBTL)`, `CD GROUND`

Noun (SCCoW), CD GROUND Noun (Gigaword), CD GROUND Noun (SUBTL), Frequency Topological Nominal (SCCoW), Frequency Topological Nominal (Gigaword), Frequency Topological Nominal (SUBTL), CD Topological Nominal (SCCoW), CD Topological Nominal (Gigaword), CD Topological Nominal (SUBTL). Previously, Inhoff and Liu (1998) and G. Yan et al. (2006) found that word frequency influences eye fixation durations on two-character words and H. M. Yang and McConkie (1999) found word frequency effects on the probability of a fixation.

H. M. Yang and McConkie (1999) furthermore observed effects of the visual complexity of a word on fixation durations and fixation probabilities. The remaining 12 predictors retrieved from the CLD are measures of the visual complexity of the GROUND noun and the topological nominal. Length GROUND Noun, Length Topological Nominal, Strokes GROUND Noun and Strokes Topological Nominal were extracted straightforwardly from the CLD. Other visual complexity measures, however, are present in the CLD at the character level only. For the number of pixels, the picture size and the number of high-level and low-level visual components, we therefore calculated word-level measures by summing the relevant character-level measures for the character in a word. This procedure resulted in 8 more word-level visual complexity measures: Pixels GROUND Noun, Pixels Topological Nominal, Picture Size GROUND Noun, Picture Size Topological Nominal, High-Level Components GROUND Noun, High-Level Components Topological Nominal, Low-Level Components GROUND Noun, Low-Level Components Topological Nominal.

We decided not to include character or sub-character level predictors as independent variables. Chapter 3 investigated character and sub-character level processing in great detail through an experimental task that was well-suited for investigating lexical processing at a small grain size: isolated word naming. The current chapter is an attempt to complement the knowledge gained in Chapter 3 with information about lexical processing at a somewhat larger grain size. Correspondingly, the experimental design was constructed to allow for an investigation of predictor effects at and above the word level, but not below the word level.

Specifically, an investigation of character or sub-character level effects in the current design is problematic due to the variation in the number of characters that GROUND nouns and topological nominals consist of. Of the 96 nouns used as grounds, 11 nouns consisted of one character, 76 nouns consisted of two characters and 9 nouns consisted of three characters. Similarly, of the 40 locative markers used in the experiment, 15 locative markers consisted of a single character, whereas 25 locative

markers consisted of two characters. The variation in the length of the GROUND nouns and the topological nominals provides a more natural language processing environment.

As a result of the variation with respect to the number of characters GROUND nouns and topological nominals consist of, a regression analysis for the full data set with lexical predictors at or below the character level would suffer from large amounts of missing data and consequently a loss of statistical power and unrepresentative results. Separate analyses for subsets of the data with the same number of characters in the GROUND noun and the topological nominal would help overcome the problem of missing data. For the current experimental data, however, the combination of three GROUND noun lengths with two topological nominal lengths would result in 6 relatively small subsets of the data. The statistical power for the analyses on each of these subsets would be limited.

The current experimental design thus does not lend itself straightforwardly for an investigation of lexical predictor effects at or below the character level. By contrast, it does provide the opportunity to investigate the effects of lexical predictors not only at, but also above the word level. In addition to the 24 word-level predictors described above, we included 9 predictors that describe properties of the prepositional construction for a total of 33 independent variables.

For each of the 384 locative phrases, we extracted the frequency and contextual diversity (defined as the number of documents a phrase occurred in) from the SCCoW and the Gigaword corpus, which resulted in four independent variables: `Phrase Frequency (SCCoW)`, `Phrase Frequency (Gigaword)`, `Phrase CD (SCCoW)` and `Phrase CD (Giga)`. No less than 140 phrases, however, did not occur in the SCCoW. Similarly, 128 phrases were not present in the Gigaword corpus. We therefore complemented the frequency and contextual diversity measures with `Phrase Frequency (Google)`, which is defined as the number of Google documents in which a phrase occurs. All phrases used in the experiment had a non-zero frequency in Google.

Next, we included two measures of the locative phrase entropy for each GROUND noun. From SCCoW and the Gigaword corpus, we extracted the phrase frequencies for all locative phrases that consist of one of the 96 GROUND nouns and one of the 40 topological nominals that occurred in the experiment. For each GROUND noun, we calculated the probability distribution over all locative phrases that contained that GROUND noun. `Entropy (SCCoW)` and `Entropy (Gigaword)` are defined as the entropy over the probability distribution for a given GROUND noun in the SCCoW and in the Gigaword corpus, respectively.

The final two construction-level predictors describe the prepositional relative entropy of a GROUND noun in the SCCoW and in the Gigaword corpus: `RE (SCCoW)` and `RE (Gigaword)`. Given estimated probabilities p (relative frequencies) of locative expressions for a given GROUND noun and estimated probabilities q of the 40 topological nominals across all 96 GROUND nouns in the experiment, relative entropy is defined as:

$$\text{Relative Entropy} = \sum_{i=1}^{96}(p_i * \log_2 (p_i/q_i)). \tag{4.2}$$

4.2.1.4 Procedure

Eye movements were recorded with an EyeLink 1000 system using a temporal resolution of 500 Hz. The experiment was run on a 17-inch CRT monitor using a 1024 by 768 pixel resolution. Participants read with the head positioned on a chin rest that was located at a distance of 70 cm from the monitor. Prior to the experiment, they were instructed to limit eye blinking to a minimum during trials. Furthermore, participants were instructed to respond as fast as possible, while retaining accuracy.

A fixation mark was shown prior to each trial. When participants fixated on this fixation mark, a locative phrase in Mandarin Chinese appeared on the screen. Participants were asked to fixate on a visual "Next" button in the bottom right corner of the screen when they completed reading the locative phrase. A fixation on this button resulted in another fixation mark. Upon fixation, this fixation mark was followed by an English translation of the locative phrase. Again, participants were asked to fixate on the visual "Next" button after reading this translation. This triggered a translation judgement screen, which contained two visual buttons located in the top half of the screen: a button on the left labeled "Correct" and a button on the right labeled "Incorrect". Participants provided translation judgements by looking at the appropriate button. Upon completion of the translation judgement task a fixation mark appeared to indicate the start of the next trial.

All text was presented in black Arial 36 point font against a white background. All fixation marks, locative phrases, and English translations were vertically centered and left-aligned at 76 pixels from the left edge of the screen. The minimum fixation duration to trigger the visual "Next" button was 500 ms, whereas the minimum fixation duration for the "Correct" and "Incorrect" buttons was 1000 ms. A 9-point grid calibration was carried out prior to the experiment, as well as after every

25 trials. The experiment had a duration of about 80 minutes, including setup, calibration and a 10 minute break halfway through the experiment.

4.2.2 Analysis

Each locative phrase consisted of three components: the preposition 在, a GROUND noun and a topological nominal. For the GROUND noun and the topological nominal we carried out three analyses: one for `Fixation Duration`, one for `Fixation Position` and one for `Fixation Probability`. Prior to analysis, all fixation durations were log-transformed to remove a rightward skew from the `Fixation Duration` distribution.

Participants did not fixate on every word in each trial. While 92.67% of all GROUND noun tokens was fixated on at least once, participants fixated on 63.00% of all preposition tokens and 34.51% of all topological nominals tokens only. Given previous fixations on the preposition and/or GROUND noun, therefore, a fixation on the topological nominal was necessary slightly more than a third of the time only. If a word was fixated on, it was typically fixated on only once. This was the case for 76.38% of all word tokens that were fixated on at least once. We therefore decided to limit our analysis of `Fixation Duration` and `Fixation Position` to first fixations. First fixations are defined as the first fixation on a word after stimulus onset.

The CLD contains 30,645 words. Nonetheless, a number of words used in the experimental items for this experiment – which was conducted prior to the creation of the CLD – are not in the CLD. First, the CLD contains one-character words and two-character words only, whereas 9 of the 96 (9.38%) GROUND nouns used in the current experiment were three character nouns. Furthermore, 24 (6.25%) experimental items contained a one-character or two-character GROUND noun or topological nominal that is not present in the CLD.

Finally, we removed the four experimental items (1.04%) that contained the GROUND noun "枕头" ("pillow"). Outside the context of the word "pillow", the second character "头" can also be a topological nominal that has a meaning similar to the English phrase "on top of" or "at the top". Given the adjacency of the second character of the GROUND noun and the topological nominal in the current stimuli, this led to confusion for the experimental items containing the GROUND noun "枕头", as reported by participants. The exclusion of experimental items prior to analysis resulted in a total data loss of 16.67%.

For the remaining experimental items, we removed individual data points for the `Fixation Position` and `Fixation Duration` analyses based on three criteria. First, fixations with a duration smaller than 50 ms or longer than 750 ms were removed from the data. This led to the exclusion of 2.18% of all fixations on prepositions, 0.86% of all fixations on GROUND nouns, and 1.34% of all fixations on topological nominals.

Second, we included first fixations on a word if and only if no previous fixations to the right of a first fixation existed. First fixations on the GROUND noun for which a previous fixation on the topological nominal existed, for instance, were excluded from the data prior to analysis to prevent information gathered through a previous fixation on the topological nominal from entering the analysis for first fixations on the GROUND noun. As a result, we removed 3.87%, 3.89% and 2.36% of the fixations on prepositions, GROUND nouns and topological nominals.

Third, to further ensure that the first fixations under investigation form a homogeneous set for which we expect lexical processing to be similar, we excluded fixations on the preposition for which a previous fixation existed (4.96%). In addition, we removed fixations on the GROUND noun (8.19%) and the topological nominal (1.52%) for which *no* previous fixations existed for the same reason. In total, the application of the three criteria for removing individual data points resulted in a data loss of 11.01% for fixations on prepositions, 12.94% for fixations on GROUND nouns and 5.22% for fixations on topological nominals.

To prevent further data loss, we excluded outliers for control variables and lexical predictors if and only if a control variable or lexical predictor was included in a reported model. Whenever relevant, we removed values for `Previous Fixation Duration` smaller than 50 (ms) and longer than 750 (ms) prior to running the model. For all other control variables and for all principal components based on lexical predictors (see below), we removed outliers further than 2.5 standard deviations from the predictor mean when a control variable or lexical predictor was included in the model.

The 33 numerical variables under investigation are highly correlated. This results in an extremely high condition number: $\kappa = 809.42$. To overcome the problem of multicollinearity we subjected the numerical predictors to a principal components analysis. Prior to this principal components analysis we applied Yeo-Johnson power transforms (Yeo & Johnson, 2000) to all predictors using version 2.1-2 of the `car` package for R (Fox & Weisberg, 2011). The resulting power-transformed predictors

were scaled prior to the principal component analysis to prevent predictors with large ranges from dominating the principal components analysis.

Consistent with our analysis of the word naming data in Chapter 3, we applied a varimax rotation to the principal components to obtain better interpretable components. The principal components analysis with varimax rotation was carried out using the `principal` function in version $1.5-8$ of the `psych` package for R (Revelle, 2015). From the set of 20 principal components, we extracted the first 11 components. Together, these 11 principal components explain 94% of the variance in the original data.

For the analyses of the `Fixation Duration`, `Fixation Position`, and `Fixation Probability`, the control variables and the 11 principal components were entered into generalized additive mixed-effect models (Hastie & Tibshirani, 1986), as implemented in version 1.8-12 of the MGCV package for R (S. Wood, 2006; S. N. Wood, 2011). GAMMs allow for the detection of non-linear effects while accounting for the variance of random effect terms such as `Participant` or `Item`. Standard GAMMs were used to model `Fixation Duration` and `Fixation Position`, whereas logistic GAMMs were used to model `Fixation Probability`.

Whenever significant, we added random intercepts for `Participant` and `Item` to a model. The effects of all other predictors and all control variables were modelled through smooth terms. To prevent uninterpretable results, we restricted all predictor smooths to 4th order non-linearities ($k = 4$). Due to the large number of models reported in this chapter, the overall pattern of results is fairly complex. In the interest of interpretability, we therefore restricted our analysis to the main effects of predictors and did not include interaction terms in the models.

The GAMs reported here provide detailed insight into the qualitative nature of predictor effects. GAMs, however, provide relatively little insight into the quantitative contribution of predictors. Therefore, we complement the GAM analyses reported here with quantitative analyses using GBMs. The results of the GBM analyses are reported in section 4.4.

4.2.3 Results

The results for Experiment 1 will be described in left-to-right order, starting with the preposition and ending with the topological nominal. A summary of the results for the lexical predictors across response variables (`Fixation Probability`, `Fixation Position`, `Fixation Duration`) and locative phrase components (preposition,

ground noun, and topological nominal) is presented in Table 4.2. The direction of an effect is indicated with + (positive relation between predictor and response variable) and − (negative relation between predictor and response variable) signs. Complex Non-linear effects are indicated with a C. For each effect, we established whether the effect was present when observations with residuals further than 2.5 standard deviations from the residual mean were removed from the model. Effects that lost significant when residual outliers were removed from the model are labelled with square brackets in Table 4.2. Furthermore, we investigated whether or not effects remained significant when random by-participants slope for the relevant predictor were added to the model. Effects that lost significance when random slopes were added to the model are labelled with round brackets in Table 4.2.

4.2.3.1 Preposition

The preposition 在 indicates the start of a locative phrase, but is semantically empty. Nonetheless, participants fixated on the preposition at least once in 63.00% of all trials. The Fixation Probability for the preposition varied between participants, as indicated by a significant random effect of Participant ($\chi^2 = 631.743$, $p < 0.001$). Furthermore, participants fixated on the preposition more often near the end of the experiment (Trial: $\chi^2 = 14.738$, $p = 0.003$). Participants decide whether or not to fixate on the preposition before any word in the locative phrase has been fixated on. Nonetheless, we also found an effect of a lexical predictor on the probability of a fixation on the preposition: PC1 ($\chi^2 = 4.673$, $p = 0.031$). The lexical predictors with the highest loadings on PC1 are CD Topological Nominal (Gigaword) (0.957), CD Topological Nominal (SCCoW) (0.955), Frequency Topological Nominal (Gigaword) (0.954) and Frequency Topological Nominal (SCCoW). The equivalent measures from SUBTLEX-CH had high loadings on PC1 as well, with loadings of 0.847 and 0.851 for CD Topological Nominal (SUBTLEX-CH) and Frequency Topological Nominal (SUBTLEX-CH), respectively.

For the topological nominal, however, frequency measures and visual complexity measures are highly correlated. As a result, the principal components analysis – even with varimax rotation – could not entirely pull apart frequency measures and visual complexity measures for the topological nominal. Although separate principal component measures exist for the visual complexity and the length of the topological nominal (see below), these measures also had high negative loadings on PC1. To be precise, we found high negative loadings for the following lexical predictors:

Table 4.2: Summary of results for Experiment 1. Plus symbols indicate a positive relation between predictor and response variable, minus symbols indicate a negative relation. The symbol C indicates a complex non-linear relation. Effects with the symbol C are discussed in more detail in the text. The number of symbols corresponds to the significance of the effects, with $p < 0.001$ for three symbols, $p < 0.01$ for two symbols and $p < 0.05$ for one symbol. Round brackets indicate that the main effect of a predictor loses significance when random by-participant slopes for that predictor are added to the model. Square brackets indicate that an effect loses significance when observations with residuals further than 2.5 standard deviations are removed from the model.

	Preposition			GROUND Noun			Topological Nominal		
	prob.	pos.	dur.	prob.	pos.	dur.	prob.	pos.	dur.
GROUND Noun									
PC8: Length				+++	+++	−−−		++	−−
PC3: Visual Complexity			−−−	+++	+++		−−−	−	
PC10: Picture Size			(−)		+++				
PC2: Frequency				−−−	−−				
PC9: Frequency (SUBTLEX-CH)									
Topological Nominal									
PC11: Length				−−−	+++		+++	+++	+++
PC7: Visual Complexity				−−	+++		+++	(++)	
PC1: Frequency	+		(+++)	+++	−−−		−−−	−−−	−−−
Construction									
PC4: Phrase Frequency		[+]			−−−		(−−−)	−	−−
PC6: Entropy					CCC				
PC5: Relative Entropy									

`Low-Level Components Topological Nominal` (−0.897), `High-Level Components Topological Nominal` (−0.896), `Picture Size Topological Nominal` (−0.877), `Pixels Topological Nominal` (−0.861), `Strokes Topological Nominal` (-0.854) and `Length Topological Nominal` (−0.797). Although we will refer to `PC1` as the frequency of the topological nominal, in a broader sense it can therefore be referred to as a measure of the ease with which the lexical representation of the topological nominal can be accessed.

As can be seen in Table 4.2, greater values of `PC1` led to a greater probability of a fixation on the preposition. Although the effect of `PC1` on the probability of a fixation is relatively subtle, it highlights two facts about lexical processing in Mandarin Chinese that we already observed in Chapter 3 and that will prove pivotal for a comprehensive understanding of locative phrase processing as well.

First, in the time between a fixation on the fixation mark (which was located at the same position as the preposition) and the first fixation on the preposition the eye of the participant was focused on the leftmost word of the locative phrase. Nonetheless, a lexical predictor related to the rightmost word in the locative phrase co-determines whether or not a preposition was fixated on. Eye movement patterns are thus influenced not only by the current word, but also by upcoming words. In Chapter 3, we found that fixations on one character allowed for pre-processing of the other character through parafoveal preview. The current findings demonstrate that parafoveal preview effects are not limited to a single character. In English 14 to 15 letters to the right of the fixation become available through parafoveal preview (Rayner, 1998). Similarly, parafoveal preview is possible for 2 to 3 characters to the right of the fixation for Chinese (Inhoff & Liu, 1997, 1998). Below, we demonstrate that parafoveal preview effects are ubiquitous for locative phrase reading in Mandarin Chinese.

Second, from an information-theoretic perspective, `PC1` can be thought of as a measure of the amount of information provided by the topological nominal. The more frequent (and the less visually complex) the topological nominal, the less information it provides. The effect of `PC1` on the probability of a fixation on the preposition therefore indicates that the smaller the amount of information provided by the topological nominal results in more fixations on the preposition. In Chapter 3 we found that a dynamic search for information characterized fixation patterns on two-character words in a word naming task. Below, we demonstrate that the allocation of resources based on the amount of information provided by the current

word and the upcoming words is a recurring theme for the current data as well. The more information-rich the current word, the more attention it receives. By contrast, upcoming words that are rich in information lead to a reduction in resources allocated to the current word.

The time between the fixation mark and the first fixation on the preposition is extremely limited. As a result, we found relatively few effects for the `Fixation Position` on the preposition. In addition to a random effect for `Participant` ($F = 47.762$, $p < 0.001$), we observed an effect of `PC4` only ($F = 4.583$, $p = 0.032$). `PC4` encodes the frequency of the locative phrase as a whole, with high loadings for `CD Phrase (SCCoW)` (0.879), `Frequency Phrase (SCCoW)` (0.870), `CD Phrase (Gigaword)` (0.829), `Frequency Phrase (Gigaword)` (0.815), and `Frequency Phrase (Google)` (0.719). More frequent phrases corresponded to more rightward fixation positions on the preposition. The effect of `PC4`, however, was not statistically robust. When residual outliers were removed from the model, the effect of `PC4` lost significance: $F = 3.526$, $p = 0.068$.

The analysis of `Fixation Duration` for fixations on the preposition yielded more interesting results. First, we observed a random effect of `Participant` ($F = 47.762$, $p < 0.001$). Furthermore, we found an effect of `X Position` ($F = 59.797$, $p < 0.001$). Fixation durations were longer for fixations near the right edge of the preposition. The effect of `X Position` is a second example of an effect that can be explained through parafoveal preview. The information that becomes available through parafoveal preview is greater for fixations near the right edge of the prepositions than for fixations near the left edge of the preposition. Therefore, upcoming words can be (pre-)processed to a greater extent during fixations near the right edge of the preposition than during fixations near the left edge of the preposition. As a result, more information is processed for more rightward fixations on the preposition, which leads to longer fixation durations.

In addition to the random effect of `Participant` and the effect of `X Position`, we observed three effects of lexical predictors on the duration of initial fixations on the preposition. First, we found two effects related to the visual complexity of the upcoming GROUND noun. `PC3` encodes the visual complexity of the GROUND noun. The lexical predictors with the highest loadings on `PC3` are `Strokes GROUND Noun` (0.957), `Low-Level Components GROUND Noun` (0.952), `High-Level Components GROUND Noun` (0.904), and `Pixels GROUND Noun` (0.901).[1] `PC10` encodes the picture

[1] `Picture Size GROUND Noun` (0.728) and `Length GROUND Noun` (0.501) also have relatively high loadings on `PC3`. These predictors, however, are encoded in independent principal components as well.

size of the GROUND noun and has a high loading for `Picture Size GROUND Noun` (0.622) only.

Both `PC3` ($F = 17.805$, $p < 0.001$) and `PC10` ($F = 2.516$, $p = 0.044$) are negatively related to fixation duration, although the main effect of `PC10` loses significance when random by-participant slopes for `PC10` are added to the model ($F = 1.381$, $p = 0.240$). The greater the amount of visual information provided by the upcoming GROUND noun, therefore, the shorter the fixation on the preposition. The effects of visual complexity thus are a second example of the fact that readers dynamically allocate resources based on the amount of information provided not only by the current word, but also by the upcoming word.

The effect of the third lexical predictor, `PC1` (i.e., the frequency of the topological nominal; $F = 14.261$, $p < 0.001$), is in line with such an interpretation of the effect of the visual complexity of the GROUND noun. The greater the frequency of the topological nominal, the longer the fixation duration of initial fixations on the preposition (see Table 4.2). When the upcoming topological nominal provides less information, the duration of the initial fixation on the preposition hence is longer. The main effect of `PC1`, however, loses significance when random by-participant slopes for `PC1` are added to the model ($F = 0.715$, $p = 0.398$) and thus shows considerable between-participants variability.

4.2.3.2 GROUND noun

While participants did not fixate on the preposition 37.00% of the time, the GROUND noun was nearly always fixated on at least once (92.67%). The `Fixation Probability` for the GROUND noun varied significantly between participants (`Participant`; $\chi^2 = 376.471$, $p < 0.001$) and items (`Item`; $\chi^2 = 40.539$, $p = 0.015$). Furthermore, we found an effect of `Trial` ($\chi^2 = 16.830$, $p < 0.001$). The effect of `Trial` was inverse U-shaped in nature, with a lower probability of fixations on the GROUND noun at the start and at the end of the experiment.

Lexical properties of the GROUND noun itself and the upcoming topological nominal also influenced the probability of a fixation on the GROUND noun. For the GROUND noun itself, our analysis revealed significant effects of three principal components: `PC8` ($\chi^2 = 66.924$, $p < 0.001$), `PC3` ($\chi^2 = 46.384$, $p < 0.001$), and `PC2` ($\chi^2 = 22.814$, $p < 0.001$). `PC8` encodes the length of the GROUND noun (highest loading: `Length GROUND Noun` (0.810); absolute values of all other loadings: < 0.30). `PC3` was introduced above and represents the visual complexity of the GROUND noun. `PC2`

encodes the frequency of the GROUND noun and has high loadings for `Frequency GROUND Noun (Gigaword)` (0.978), `Frequency GROUND Noun (SCCoW)` (0.977), `CD GROUND Noun (Gigaword)` (0.967) and `CD GROUND Noun (SCCoW)` (0.977).[2]

The measures of the visual complexity of the GROUND noun, `PC8` and `PC3` had a positive relation with fixation probability: more complex GROUND nouns were fixated on more often. By contrast, `PC2` had a negative relation with fixation probability: more frequent GROUND nouns were fixated on less often. These findings fit well with the idea that eye movement patterns are determined by the amount of information provided by the current word. The more information a GROUND noun provides, the greater the probability of a fixation on that GROUND noun.

Lexical properties of the upcoming topological nominal likewise influenced the probability of a fixation on the GROUND noun. `PC11` and `PC7` encode the length and the visual complexity of the topological nominal, respectively. Due to the fact that the principal components analysis had a hard time separating visual complexity and frequency for the topological nominal, the highest loadings on `PC11` and `PC7` are not as high as the highest loadings for the other principal components. The only predictor with a relatively high loading on `PC11` is `Length Topological Nominal` (loading: 0.550; absolute values of all other loadings: < 0.30). `Strokes Topological Nominal` (0.452), `Low-Level Components Topological Nominal` (0.385), `Pixels Topological Nominal` (0.360), and `High-Level Components Topological Nominal` (0.331) are the lexical predictors with the highest loadings on `PC7`. Unfortunately, `CD Topological Nominal (SUBTLEX-CH)` (0.428) and `Frequency Topological Nominal (SUBTLEX-CH)` (0.341) also had fairly high loadings on `PC7`. Nonetheless, as will become clear below, `PC7` behaves much more like a visual complexity measure than like a frequency measure in the analyses reported here.

The analysis of the probability of a fixation on the GROUND noun showed similar effects of `PC11` ($\chi^2 = 13.198$, $p < 0.001$) and `PC7`. ($\chi^2 = 7.519$, $p = 0.006$). The probability of a fixation on the GROUND noun is lower when the topological nominal is a visually complex two-character word than when it is a visually simple single-character word. We furthermore observed an effect of `PC1` (i.e., the frequency of the topological nominal; $\chi^2 = 13.198$, $p < 0.001$). As can be seen in Table 4.2, more frequent topological nominals resulted in more fixations on the GROUND noun. Again,

[2]The frequency (`Frequency GROUND Noun (SUBTLEX-CH)`, loading: 0.742) and contextual diversity (`CD GROUND Noun (SUBTLEX-CH)`, loading: 0.737) of the GROUND noun in SUBTLEX-CH have high loadings on `PC2` as well. The frequency of the GROUND noun in SUBTLEX-CH, however, is also encoded in a separate principal component: `PC9`.

these effects suggest that participants dynamically allocate their resources based on the information-richness of the different words in a locative phrase. The more information the upcoming topological nominal provides, the lower the probability of a fixation on the GROUND noun.

Before fixating on the GROUND noun, participants decide what the optimal fixation position would be. Due to parafoveal preview during the fixation on the preposition, a lot of information is available to inform this decision. As a result, a wide range of lexical predictors co-determined the `Fixation Position` of the initial fixation on the GROUND noun. Before discussing these effects, however, we briefly discuss the effects of the control variables on the position of a fixation.

We observed random effects of both `Participant` (`Participant`: $F = 177.010$, $p < 0.001$) and `Item` ($F = 0.534$, $p < 0.001$). In addition, we found effects of three further control variables: `Previous Fixation Duration` ($F = 121.614$, $p < 0.001$), `Partial Saccade Length` ($F = 938.628$, $p < 0.001$) and `Trial` ($F = 4.115$, $p = 0.004$). The effect of `Previous Fixation Duration` was characterized by a complicated non-linear pattern. For the part of the predictor range with most data points (i.e., the middle of the predictor range), however, longer previous fixation durations led to fixations that were less far into the GROUND noun. Similarly, previous fixations that were less far into the preposition result in more leftward initial fixation positions on the GROUND noun. Finally, fixation positions were somewhat further into the GROUND noun at the start of the experiment than at the end of the experiment.

Lexical properties of the GROUND noun itself, the topological nominal and the locative phrase as a whole co-determined fixation positions on the GROUND noun. Fixations were further into GROUND nouns that were more visually complex, as indicated by strong effects of `PC8` (i.e., the length of the GROUND noun; $F = 174.608$, $p < 0.001$), `PC3` (i.e., the visual complexity of the GROUND noun; $F = 27.071$, $p < 0.001$), and `PC10` (i.e., the picture size of the GROUND noun; $F = 19.128$, $p < 0.001$). By contrast, the effect of `PC2` (i.e., the frequency of the GROUND noun; $F = 19.128$, $p < 0.001$) indicated that fixations were less far into the GROUND noun for high frequency GROUND nouns as compared to low frequency GROUND nouns.

The fixation position analysis revealed similar effects of lexical properties of the topological nominal. The effects of `PC11` (i.e., the length of the topological nominal; $F = 114.506$, $p < 0.001$) and `PC7` (i.e., the visual complexity of the topological nominal; $F = 77.203$, $p < 0.001$) indicate that a greater visual complexity of

the topological nominal resulted in more rightward fixations on the GROUND noun. Again, the effect of frequency was in the opposite direction: fixations were less far into the GROUND noun when the upcoming topological nominal was more frequent (`PC1`: $F = 426.392$, $p < 0.001$).

The results for lexical properties of the GROUND noun and topological nominal fit well with the hypothesis that readers continuously adapt their eye movement patterns based on information-theoretic properties of the locative phrase. When deciding where to fixate on the GROUND noun, the amount of information provided by the GROUND noun and the topological nominal determines the optimal fixation position. The richer both words are in information, the more rightward the optimal viewing position. In other words: resources are allocated to the part of the locative phrase that provides most information. The effect of `PC4` (i.e., the frequency of the locative phrase as a whole; $F = 107.037$, $p < 0.001$) is in line with such an interpretation of the effects of lexical properties of the topological nominal and the GROUND noun on the position of the initial fixation on the GROUND noun: fixations are less far into the GROUND noun for more frequent locative phrases.

In addition to the effect of `PC4`, we found a significant effect of another principal component at the construction level: `PC6` ($F = 6.317$, $p < 0.001$). The lexical predictors with the highest loadings on `PC6` are `Entropy (Gigaword)` (0.968) and `Entropy (SCCoW)` (0.833). The absolute values of the loadings for all other lexical predictors are smaller than 0.30. The effect of `PC6` has a complex non-linear nature and is therefore presented in Figure 4.1.

As can be seen in Figure 4.1, the overall trend of the effect is downward, with fixations being less far into the GROUND noun for higher values of `Entropy`. How-

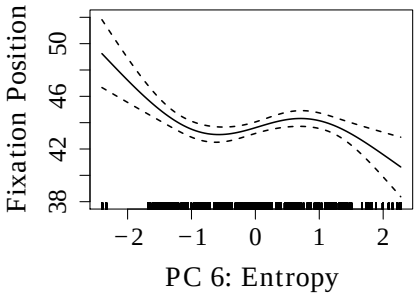

Figure 4.1: Fixation Position results for locative phrases presented in isolation: GROUND noun. Effect of PC6: Entropy.

ever, for predictor values between -1 and 1 the effect reverses. A majority of the locative phrases (61.56%) have values of PC6 that are between -1 and 1. For these locative phrases a higher entropy leads to fixations that are somewhat further into the GROUND noun. Uncertainty about the exact nature of the entropy effect on the position of the initial fixation on the GROUND noun thus remains.

The position of a fixation for initial fixations on the GROUND noun was influenced by a large number of lexical predictors. The model for Fixation Duration, by contrast, revealed an effect of PC8 (i.e., the length of the GROUND noun; $F = 11.761$, $p < 0.001$) only. Fixation durations were shorter for two-character GROUND nouns as compared to one-character GROUND nouns. The nature of the effect of PC8 is surprising. Based on what we have seen thus far, we would expect the increased amount of information provided by two-character GROUND nouns as compared to one-character GROUND nouns to result in longer fixation durations. The effects for fixation duration of initial fixations on the topological nominal reported below are in line with this expectation.

One potential explanation for the effect of PC8 is that two-character GROUND nouns tend to have more specific meanings than one-character GROUND nouns. Two-character GROUND nouns therefore provide stronger expectations about the upcoming topological nominal. On average, the one-character GROUND nouns in our experiment appear with 15.27 different nominals in locative phrases in the SCCoW, whereas the two-character GROUND nouns in our experiment appear with 9.48 different nominals in locative phrases in the SCCoW ($t(11.23) = 3.012$, $p = 0.011$). Consequently, parafoveal pre-processing and fixation planning for the upcoming topological nominal may be more efficient for two-character GROUND nouns as compared to one-character GROUND nouns, which would result in shorter fixation durations.

In addition to the effect of PC8, we observed random effects of Participant ($F = 104.955$, $p < 0.001$) and Item ($F = 0.267$, $p = 0.001$), as well as effects of a number of control variables. Fixation durations were shorter for trials that occurred later in the experiment (Trial: $F = 135.663$, $p < 0.001$), for more rightward fixations (X Position: $F = 10.211$, $p < 0.001$), and for fixations for which the previous fixation duration was longer (Previous Fixation Duration: $F = 7.061$, $p = 0.008$). By contrast, fixation durations were longer when the size of the incoming saccade was larger (Saccade Length: $F = 4.612$, $p = 0.032$) and when the vertical fixation position was higher (Y Position: $F = 3.862$, $p = 0.049$). However, the support for the effects of Saccade Length and Y Position is relatively weak.

4.2.3.3 Topological Nominal

The topological nominal is the third and final word in a locative phrase. Whereas GROUND nouns were almost always fixated on at least once, only 34.51 % of all topological nominals were fixated on. The low probability of a fixation on the topological nominal is consistent with previous findings by H. C. Chen et al. (2003), who reported that the proportion of words that are not fixated on (i.e., that are "skipped") is much higher in Chinese than in English.

The `Fixation Probability` for the topological nominal varied between participants (`Participant`: $\chi^2 = 844.655$, $p < 0.001$) and items (`Item`: $\chi^2 = 79.506$, $p < 0.001$). However, the effect of `Item` lost significance when observations with residuals further than 2.5 standard deviations from the residual mean were removed from the model ($F = 0.105$, $p = 0.060$). Furthermore, the probability of a fixation on the topological nominal was higher at the start of the experiment (`Trial`; $\chi^2 = 73.191$, $p < 0.001$). This suggests that participants optimize their reading strategy throughout the experiment, such that there is a decreased need for fixations on the topological nominal near the end of the experiment.

Unsurprisingly, the probability of a fixation on the topological nominal was primarily influenced by lexical properties of the topological nominal itself. The effects of lexical properties of the topological nominal on the fixation probability for the topological nominal closely resemble the effects of lexical properties of the GROUND noun on the fixation probability for the GROUND noun. The probability of a fixation on the topological nominal is greater for visually complex two-character topological nominals than for visually simple single-character topological nominals, as indicated by strong effects of `PC11` (i.e., the length of the topological nominal; $\chi^2 = 102.773$, $p < 0.001$) and `PC7` (i.e., the visual complexity of the topological nominal; $\chi^2 = 40.259$, $p < 0.001$). In addition, the probability of a fixation on the topological nominal is smaller for more frequent topological nominals as compared to less frequent topological nominals (`PC1`: $\chi^2 = 880.466$, $p < 0.001$).

We furthermore observed an effect of a lexical property of the construction as a whole. The relation between `PC4` (i.e., phrase frequency) and fixation probability was negative ($\chi^2 = 122.627$, $p < 0.001$). The topological nominal was thus fixated on less often in more frequent locative phrases. Note, however, that the effect of `PC4` lost significance when random by-participant slopes for `PC4` were added to the model ($\chi^2 = 3.678$, $p = 0.056$). The effect of phrase frequency thus varied considerably between participants. Nonetheless, the results for the lexical properties of the topological

nominal and the construction as a whole are in line with the claim that eye fixation patterns are to a large extent determined by information-theoretic demands. The more frequent the topological nominal and the phrase as a whole, the less resources need to be allocated to the topological nominal.

The final effect of a lexical predictor on the probability of a fixation was an effect of a lexical property of the GROUND noun: `PC3` (i.e., the complexity of the GROUND noun; $\chi^2 = 34.368$, $p < 0.001$). Interestingly, visually complex GROUND nouns elicited fewer fixations on the topological nominal than visually simple GROUND nouns. Recall that the probability of a fixation on the GROUND noun was greater for more complex GROUND nouns. Furthermore, the fixation position for visually complex GROUND nouns was more rightward as compared to the fixation position for less complex GROUND nouns. The increased probability of a fixation on the GROUND noun and the more rightward fixation position on the GROUND noun may have allowed for more parafoveal preview while fixating on visually complex GROUND nouns as compared to visually simple GROUND nouns and thus for fewer fixations on the topological nominal, thus reducing the need for additional fixations on the nominal.

When a topological nominal was fixated on, the `Fixation Position` of the initial fixation was co-determined by a number of control variables. First, the position of a fixation varied between participants (`Participant`: $F = 19.313$, $p < 0.001$) and – to a lesser degree – items (`Item`: $F = 0.121$, $p = 0.038$). In addition, we observed significant effects for `Partial Saccade Length` ($F = 246.763$, $p < 0.001$) and `Previous Fixation Duration` ($F = 7.984$, $p < 0.001$). As was the case for fixation positions on the GROUND noun, fixations were further into the topological nominal when previous fixations were shorter and closer to the topological nominal. The effect of `Trial` ($F = 13.139$, $p < 0.001$) was in the opposite direction of the effect of `Trial` observed for the position of the initial fixation on the GROUND noun: fixation positions are further into the topological nominal near the end of the experiment. Finally, we observed an effect of `Y Position` ($F = 8.505$, $p < 0.001$), with fixations that are lower on the screen being further into the topological nominal.

In addition to the control variables, we also found effects of a number of lexical predictors on the position of the initial fixation on the topological nominal. First, the length (`PC11`; $F = 54.087$, $p < 0.001$), the visual complexity (`PC7`; $F = 5.064$ p

$= 0.004$)[3] and the frequency (PC1; $F = 33.734$, $p < 0.001$) of the topological nominal showed effects that were highly similar to the corresponding measures of the GROUND noun in the model for fixation position for the GROUND noun. Fixations were more rightward for two-character words than for one-character words, particularly when these words were visually complex and infrequent. We furthermore observed an effect of phrase frequency (PC4: $F = 5.064$, $p = 0.025$). This effect was in line with the results for fixation position on the GROUND noun as well: fixations positions on the topological nominal were further into the topological nominal for less frequent phrases.

Lexical predictors related to the GROUND noun also co-determined fixation positions on the topological nominal. Longer GROUND nouns resulted in more rightward fixation positions on the topological nominal (PC8: $F = 5.107$, $p = 0.002$). This effect of the length of the GROUND noun is consistent with the effects reported above: longer words – including preceding and upcoming words – yield more rightward fixation positions.

The fixation position analysis furthermore revealed an effect of PC3, the visual complexity of the GROUND noun ($F = 3.521$, $p = 0.033$). Fixations were less far into the topological nominal if the preceding word was visually complex. One interpretation of this effect of PC3 is that the increased number of fixations on the GROUND noun and the more rightward fixation position on the GROUND noun for visually complex GROUND nouns resulted in increased parafoveal preview of the topological nominal. Under this hypothesis, the more leftward fixation position on the topological nominal reflects additional preprocessing of the topological nominal for locative phrases with visually complex GROUND nouns.

The third response variable, Fixation Duration, revealed a random effect of Participant ($F = 17.946$, $p < 0.001$), as well as effects of four other control variables. Fixation durations were shorter for fixations that were less far into the word (X Position: $F = 143.772$, $p < 0.001$). Fixations durations were also longer for fixations that were lower on the screen (Y Position: $F = 10.091$, $p = 0.002$) and for fixations for which the duration of the previous fixation was longer (Previous Fixation Duration: $F = 2.994$, $p = 0.021$). The effects of Y Position and Previous Fixation Duration thus are opposite to the effects of Y Position and Previous Fixation Duration on Fixation Duration for fixations on the GROUND noun.

[3]The effect of PC7 lost significance when by-participant random slopes for PC7 were added to the model ($F = 3.400$, $p = 0.065$). The effect of PC7 thus shows considerable between-participant variation.

Finally, we observed an effect of `Trial` ($F = 23.797$, $p < 0.001$). Fixation durations were shorter near the end of the experiment. Again, this suggests that readers optimize their reading strategy through the experiment, such that near the end of the experiment fewer resources are needed to successfully process the topological nominal.

We furthermore observed effects of lexical properties of the topological nominal itself, the preceding GROUND noun and the construction as a whole on fixation durations on the topological nominal. Longer preceding GROUND nouns afforded shorter fixation durations on the topological nominal (`PC8`: $F = 4.386$, $p = 0.005$). Recall that the probability of a fixation on the preceding GROUND noun was greater for two-character words as compared to one-character words. Furthermore, fixations on two-character GROUND nouns were further into the word than fixations on one-character GROUND nouns. Again, therefore, the effect of `PC8` may be a result of increased parafoveal preview of the topological nominal when fixating on preceding two-character GROUND nouns as compared to preceding one-character GROUND nouns.

In addition to the effect of the length of the GROUND noun, we also observed an effect of the length of the topological nominal (`PC11`: $F = 21.715$, $p < 0.001$). As expected, fixation durations were longer for two-character topological nominals as compared to one-character topological nominals. Furthermore, fixation durations were shorter for more frequent topological nominals than for less frequent topological nominals, as indicated by an effect of `PC1` ($F = 106.649$, $p < 0.001$). This finding is in line with G. Yan et al. (2006), who found that fixation durations tend to be shorter for more frequent words.

Finally, as was the case for the probability of a fixation on the topological nominal and the position of such a fixation, we observed an effect of `PC4` (i.e., phrase frequency; ($F = 8.615$, $p = 0.003$) on `Fixation Duration`. The topological nominal was fixated on less long for more frequent phrases. Consistent with our interpretation of the results thus far, the effects of lexical properties of the topological nominal and the construction as a whole demonstrate that words or phrases that provide a lot of information require additional resources and, as a result, extra processing time.

4.2.4 Discussion

Experiment 1 investigated locative phrase reading for locative phrases presented in isolation. During the experiment, we measured the eye movements of the participants. For each of the three components of a locative phrase – the preposition 在, a GROUND noun and a topological nominal – we extracted the probability of a fixation on that component, as well as the position and duration of the first fixation. The results of a principal components regression using GAMMs revealed a number of interesting insights into lexical processing above the word level in Mandarin Chinese.

First, information processing is not limited to the word that is fixated on. Instead, eye movement patterns are characterized by a substantial amount of joint processing. Lexical properties of the GROUND noun and, albeit more subtly, the topological nominal start influencing eye movement patterns during or even prior to fixations on the preposition. Lexical properties of the topological nominal and the locative phrase as a whole co-determine the position of the initial fixation on the GROUND noun. Predictors describing lexical properties of the topological nominal furthermore influence the probability of a fixation on the GROUND noun. Parafoveal preview thus allows for widespread pre-processing of words before these words are fixated on.

Second, a driving force behind fixation patterns on locative phrases is a dynamic search for information. Fixation durations on the preposition are shorter when the upcoming GROUND noun and topological nominal provide more information. Similarly, the probability of a fixation on the GROUND noun is greater when the GROUND noun provides more information, but smaller when the upcoming topological nominal provides more information. Conversely, a fixation on the topological nominal is more likely when the topological nominal provides a lot of information, but less likely when the preceding GROUND noun provides a lot of information. Furthermore, fixations are further into individual words (and thus into the locative phrase as a whole) when the rest of the phrase provides more information. We conclude, therefore, that resources are dynamically allocated based on the amount of information provided by the word that is fixated on, by the preceding word, by the upcoming word, and by the locative phrase as a whole. This dynamic allocation of resources leads to highly efficient processing strategy, due to which 65.49% of all topological nominal tokens do not require a fixation at all (c.f., H. C. Chen et al., 2003).

Third, in addition to the expected effects of the frequency (see e.g., Inhoff & Liu, 1998; H. M. Yang & McConkie, 1999; G. Yan et al., 2006) and visual complexity (see H. M. Yang & McConkie, 1999) of the GROUND noun and the topological nominal, the analysis of the data for Experiment 1 revealed frequency effects of the locative phrase as a whole and the entropy of the locative phrase on the fixation patterns on the GROUND noun and the topological nominal.

Fixation positions were further into the GROUND noun and the topological nominal for frequent locative phrases as compared to less frequent locative phrases. Furthermore, fixation durations on the topological nominal were shorter for more frequent locative phrases. The results for phrase frequency are consistent with previous findings by Rayner et al. (2005), who found more fixations and longer fixation durations for words with a lower subjective predictability rating in a sentence reading task. They also fit well with the (forward and backward) transitional probability effects on eye fixation durations in a sentence reading task reported by H. C. Wang et al. (2010). The nature of the current phrase frequency effects suggests that the language processing system is sensitive to co-occurrence patterns above the word level and uses this information to process locative phrases in an optimal manner.

The results for Experiment 1 indicate that locative phrase processing is a highly dynamic process characterized by a continuous search for the locus of information based on information obtained through both foveal and parafoveal vision. The locative phrases in Experiment 1, however, are presented in isolation. In real-life language processing locative phrases are rarely presented in isolation. Hence, we carried out an additional experiment in which we present locative phrases in sentence contexts. Below, we describe the design and results of this experiment.

4.3 Experiment 2

4.3.1 Methods

4.3.1.1 Participants

Thirty participants took part in the experiment. All participants were native readers of Mandarin Chinese born in mainland China and living in Tübingen. Their mean age was 25.55 (sd: 4.45). Twenty-four participants were female, seven were male. All participants had normal or corrected to normal vision. Participants received 20 euro for their participation.

4.3.1.2 Materials

We embedded the 384 locative phrases from Experiment 1 in sentences containing an AGENT noun, a FIGURE noun and a verb. The locative phrases were embedded in two types of sentence, which differed with respect to the position of the locative phrases within the sentence. We refer to the two types of sentences as `Early Locative` and `Late Locative` sentences. `Early Locative` sentences were of the following structure:

人凤	在	书	中	粘	一	张	贴纸
Renfeng	PREP	book	in(side)	glue	one	CLASSIFIER	sticker

"Renfeng glued a sticker in the book"

Each `Early Locative` sentence started with a two-character proper noun that has the semantic role AGENT. This proper noun is followed by one of the locative phrases from Experiment 1. The next word in the sentence is a verb. `Early Locative` sentences end with a FIGURE noun phrase, which consisted of a numeral, a classifier and a noun. Numerals describe quantitative properties of the FIGURE noun (e.g., 一 ("one"), 二 ("two")), whereas classifiers describe qualitative properties (e.g., 张 refers to objects that have flat surfaces, such as "paper", "table" or "photo"; 双 refers to objects that form pairs, such as "shoes", "chopsticks", "eyes").

In `Late Locative` sentences the FIGURE noun and the verb preceded the locative phrase:

叶军	把	一	个	纸箱	留	在	街道	上
Ye Jun	OBJ. MARKER	one	CLASSIFIER	carton	left	PREP	street	on

"Ye Jun left a carton on the street"

Similar to `Early Locative` sentences, each `Late Locative` sentence started with a two-character proper noun that serves as AGENT. An object marker, indicating that the upcoming noun is a direct object, followed the proper noun. In the sentences in this experiment this direct object was a FIGURE noun. As in the `Early Locative` sentences, a numeral and a classifier preceded the FIGURE noun. The FIGURE noun phrase was followed by a verb. The sentence concluded with one of the locative phrases from Experiment 1.

The materials for Experiment 2 consisted of 192 `Early Locative` and 192 `Late Locative` sentences. A total of 96 unique verbs were used in the experiment: 85 one character verbs and 11 two-character verbs. Similarly, a set of 96 unique FIGURE

nouns was used in the experiment: 14 one-character nouns and 82 two-character nouns. Each verb, FIGURE and GROUND was repeated four times in the experiment. To prevent within-experiment priming, all verb-FIGURE, verb-GROUND and FIGURE-GROUND combinations in the experiment were unique.

Analogous to Experiment 1 the experimental task was translation verification. We translated each `Early Locative` and `Late Locative` sentence to English. Fifty percent of the translations were correct, the other fifty percent were incorrect. Incorrect translations differed from the sentences in Mandarin with respect to either the verb, the FIGURE, the GROUND or the topological nominal. Translation errors were evenly distributed across these translation error types.

4.3.1.3 Design

The experimental design was identical to the design for Experiment 1.

4.3.1.4 Procedure

The experimental procedure was similar to the experimental procedure for Experiment 1, with two notable changes. First, we reduced the font size for all text from 36 point to 28 point to ensure that all sentences would fit on a single line. Second, due to a change in the setup of the experimental lab, stimuli were presented on a 25.9 inch flatscreen monitor rather than on the 17 inch LCD monitor used in Experiment 1. As a result of these changes, the number of characters per degree of visual angle was slightly reduced in Experiment 2 (0.66) as compared to Experiment 1 (0.75; difference: 14%). The increased length of the stimuli increased the average duration of the experiment from 80 to 105 minutes.

4.3.2 Analysis

As was the case for Experiment 1, we carried out three analyses for each of the three components in a locative phrase: a `Fixation Duration` analysis, a `Fixation Position` analysis and a `Fixation Probability` analysis. We conducted separate analyses for `Early Locative` sentences and `Late Locative` sentences. For `Early Locative` sentences, 92.28% of all GROUND noun tokens, 58.72% of all preposition tokens and 80.86% of all topological nominal tokens were fixated on at least once. For `Late Locative` sentences the corresponding percentages were substantially smaller at 87.76%, 51.90% and 38.25%.

Prior to analysis we removed stimuli with locative phrases that contained a topological nominal or GROUND noun that was not in the CLD. Furthermore, we again removed the four experimental items that contained the GROUND noun 枕头 ("pillow"). For the remaining experimental items, we removed individual data points for the `Fixation Position` and `Fixation Duration` analyses based on the same three criteria used in the analysis of the data for Experiment 1. First, we removed 1.08%, 0.70%, and 1.01% of the data for the GROUND noun, the preposition and the topological nominal as fixation duration outliers.

Second, we removed initial fixations on the GROUND noun, preposition and topological nominal for which previous fixations to the right existed. As a result, we removed 7.64%, 23.64% and 10.06% of the fixations on GROUND nouns, prepositions and topological nominals. For the preposition, this procedure resulted in substantial data loss (23.64%). We decided to nonetheless apply this criterion to ensure that the set of initial fixations on the preposition is homogeneous and that the results for Experiment 2 can readily be compared to the results for Experiment 1.

Third, we excluded fixations for which no previous fixation existed. This led to the exclusion of 0.59% of the data for GROUND nouns, 1.56% of the data for prepositions and 0.09% of the data for topological nominals. The application of the three criteria for removing individual data points resulted in a total data loss of 9.30% for fixations on prepositions, 25.90% for fixations on GROUND nouns and 11.61% for fixations on topological nominals. As before, outliers for individual predictors were removed if and only if a control variable or lexical predictor was included in a reported model.

The same set of principal components was used for the analyses of the experimental data from Experiment 1 and Experiment 2. Again, we analyzed the data using GAMMs, with random effects for `Participant` and `Item` and smooth terms limited to 4-th order non-linearities for all other predictors. The results of analogous analyses of the quantitative contribution of predictor using GBMs are reported in section 4.4.

4.3.3 Results Early Locative sentences

A summary of the results for the `Early Locative` sentences is presented in Table 4.3. The probability of a fixation on the preposition and the GROUND noun was similar for the `Early Locative` sentences (preposition: 58.72%; GROUND noun: 92.28%) and the locative phrases presented in isolation in Experiment 1 (prepo-

Table 4.3: Summary of results for *Early Locative* sentences in Experiment 2. Plus symbols indicate a positive relation between predictor and response variable, minus symbols indicate a negative relation. The symbol C indicates a complex non-linear relation. Effects with the symbol C are discussed in more detail in the text. The number of symbols corresponds to the significance of the effects, with $p < 0.001$ for three symbols, $p < 0.01$ for two symbols and $p < 0.05$ for one symbol. Round brackets indicate that the main effect of a predictor loses significance when random by-participant slopes for that predictor are added to the model. Square brackets indicate that an effect loses significance when observations with residuals further than 2.5 standard deviations are removed from the model.

	Preposition			GROUND Noun			Topological Nominal		
	prob.	pos	dur.	prob.	pos.	dur.	prob.	pos.	dur.
GROUND Noun									
PC8: Length	−		− − −	+ + +	++	− − −	−	+ + +	
PC3: Visual Complexity				+ + +		C	−−		− − −
PC10: Picture Size				[+]				−	
PC2: Frequency				− − −	+				
PC9: Frequency (SUBTLEX-CH)				− − −				+ + +	
Nominal									
PC11: Length			(+)				+ + +	+ + +	− − −
PC7: Visual Complexity					+			+	
PC1: Frequency							− − −	− − −	− − −
Construction									
PC4: Phrase Frequency							− − −	− − −	
PC6: Entropy							+		
PC5: Relative Entropy					+				

sition: 63.00%; GROUND noun: 92.67%). However, whereas topological nominals were fixated on only 34.51% of the time in Experiment 1, 80.86% of all topological nominal tokens was fixated on in `Early Locative` sentences. When a topological nominal is not the final word of the input, therefore, it is fixated on much more often than when it is the final word of the input.

4.3.3.1 Preposition

The `Fixation Probability` for the preposition differed significantly between participants (`Participant`: $\chi^2 = 424.409$, $p < 0.001$) and items (`Item`: $\chi^2 = 50.200$, $p = 0.004$). Furthermore, the preposition was fixated on less often near the end of the experiment (`Trial`: $\chi^2 = 19.121$, $p < 0.001$). The effect of `Trial` thus was opposite to the effect of trial on the probability of a fixation on the preposition in Experiment 1. This suggests that in the context of a sentence participants learn to fixate less on the semantically empty preposition throughout the experiment.

The analysis of the probability of a fixation furthermore revealed an effect of the length of the GROUND noun (`PC8`: $\chi^2 = 4.778$, $p = 0.029$). Prepositions were fixated on less often when the upcoming GROUND noun was longer. This finding fits well with the results for Experiment 1 and demonstrates that both for locative phrases presented in isolation and for locative phrases presented in sentential contexts fixation patterns are co-determined by the amount of information provided by upcoming words. The more information-rich the upcoming word, the lower the probability of a fixation on the current word.

Lexical predictors did not influence `Fixation Position` for the preposition. We did, however, observe a random effect of `Participant` ($F = 3.192$, $p < 0.001$), as well as significant effects of the control variables `Partial Saccade Length` ($F = 88.462$, $p < 0.001$) and `Trial` ($F = 3.940$, $p = 0.016$). Fixations were further into the preposition when the previous fixation on the sentence-initial proper noun was closer to the left edge of the preposition and near the end of the experiment.

The `Fixation Duration` model for initial fixations on the preposition similarly revealed a random effect of `Participant` ($F = 17.872$, $p < 0.001$). Furthermore, `Saccade Length` ($F = 17.691$, $p < 0.001$) co-determined not only the position of a fixation, but also its duration: fixation durations were shorter when incoming saccade sizes were smaller. The final control variable that showed an effect on fixation durations was `Previous Fixation Duration` ($F = 4.317$, $p = 0.038$): longer fixations on the preceding proper noun resulted in longer fixations on the preposition.

We observed no effects of lexical predictors on the position of the initial fixation on the preposition. By contrast, the length of both the GROUND noun (PC8: $F = 5.939$, $p < 0.001$) and the topological nominal (PC11: $F = 3.327$, $p = 0.021$) had an effect on the duration of such fixations. As we had come to expect, fixation durations on the preposition were shorter when the upcoming GROUND noun consisted of two characters as compared to when it consisted of a single character only. Even when locative phrases are embedded in sentential contexts, therefore, fixation patterns are indicative of a dynamic search for information.

Surprisingly, the effect of PC11 was in the opposite direction: fixation durations on the preposition were longer when the upcoming GROUND noun consisted of two characters. The effect of PC11, however, loses significance when random by-participant slopes for PC11 are added to the model ($F = 2.451$, $p = 0.067$). Therefore, the main effect of PC11 is not statistically robust. This suggests that participants develop different reading strategies throughout the experiment.

4.3.3.2 GROUND noun

The model for the Fixation Probability on the GROUND noun revealed significant variation between participants (Participant: $\chi^2 = 301.512$, $p < 0.001$) and items (Item: $\chi^2 = 32.831$, $p = 0.019$). Furthermore, the probability of a fixation on the GROUND noun was influenced by all principal components that encode lexical properties of the GROUND noun. As expected, two-character (PC8: $\chi^2 = 26.124$, $p < 0.001$), visually complex (PC3: $\chi^2 = 30.742$, $p < 0.001$) GROUND nouns with a greater picture size (PC10: $\chi^2 = 4.511$, $p = 0.034$) were fixated on more often than single-character visually simple GROUND nouns with a smaller picture size. Note, however, that the effect of PC10 lost significance when observations with residuals further than 2.5 from the residual mean were removed from the model ($\chi^2 = 3.351$, $p = 0.067$).

Furthermore, more frequent GROUND nouns were fixated on less often, as indicated by effects of the principal components encoding the frequency of the GROUND noun in the SCCoW and in the Gigaword corpus (PC2: $\chi^2 = 21.142$, $p < 0.001$) and the frequency of the GROUND noun in SUBTLEX-CH (PC9 (highest loadings: CD GROUND Noun (SUBTLEX-CH) (0.612), Frequency GROUND Noun (SUBTLEX-CH) (0.593), absolute values of all other loadings < 0.30): $\chi^2 = 13.159$, $p < 0.001$). GROUND nouns that provide more information therefore are fixated on more often than GROUND nouns that provide less information.

The Fixation Position of initial fixations on the GROUND noun also differed significantly between participants (Participant: $F = 18.843$, $p < 0.001$). In addition, fixations were further into the GROUND noun when the previous fixation was shorter (Previous Fixation Duration: $F = 9.248$, $p = 0.002$) and closer to the left edge of the GROUND noun (Partial Saccade Length: $F = 404.073$, $p < 0.001$).

Consistent with the results for Experiment 1, fixations were further into the GROUND noun when the GROUND noun itself was longer (PC8: $F = 5.293$, $p = 0.002$) and when the topological nominal was visually more complex (PC7: $F = 5.035$, $p = 0.026$). As was the case in Experiment 1, therefore, fixation positions on the GROUND noun reflect the amount of information provided by both the GROUND noun itself and by the upcoming topological nominal.

Based on these findings and the results for Experiment 1, we would expect low frequency GROUND nouns to lead to more rightward fixations as well. Surprisingly, however, the frequency of the GROUND noun showed a positive relation with fixation position (PC2: $F = 6.554$, $p = 0.010$). A straightforward explanation for the positive relation between PC2 and fixation position is not available. Therefore, it will be interesting to see if this effect of the frequency of the GROUND noun on the position of a fixation is observed for Late Locative sentences as well.

A final lexical predictor effect on the position of the initial fixation on the GROUND noun is the effect of PC5 ($F = 3.903$, $p = 0.048$). The lexical predictors with the highest loadings on PC5 are RE (SCCoW) (0.952) and RE (Gigaword) (0.936). Absolute values of all other loadings on PC5 are smaller than 0.30. PC5 thus encodes the relative entropy of the GROUND noun. The frequency distribution of topological nominals for GROUND nouns with high relative entropy is less similar to the frequency distribution of topological nominals across all other GROUND nouns than the frequency distribution of topological nominals for GROUND nouns with low relative entropy. The amount of uncertainty about the identity of the topological nominal – and therefore the amount of information provided by the topological nominal – thus is greater for high values of PC5 than for low values of PC5. Unsurprisingly, therefore, we observed a positive relation between PC5 and fixation position: fixations are further into the GROUND noun for GROUND nouns with high relative entropy. However, given the p-value of 0.048, the support for the effect of PC5 is rather weak.

The `Fixation Duration` analysis revealed a random effect of `Participant` ($F = 29.299$, $p < 0.001$) and an effect of `Trial` ($F = 5.382$, $p < 0.001$). Fixation durations on the GROUND noun were shorter at the start of the experiment. As was the case for the effect of `Trial` on the fixation probability for the preposition, therefore, the effect of `Trial` on the duration of the initial fixation on the GROUND noun is in the opposite direction of the equivalent effect in Experiment 1.

The effect of `PC8` (i.e., the length of the GROUND noun; $F = 13.472$, $p < 0.001$), by contrast, was qualitatively similar to the effect of `PC8` on fixation durations on the GROUND noun in Experiment 1: fixation durations were shorter for two-character GROUND nouns as compared to one-character GROUND nouns. We furthermore found an effect of `PC3` (i.e., the visual complexity of the GROUND noun; $F = 5.326$, $p = 0.001$). The effect of `PC3` had a complex non-linear nature. For clarity, this effect is therefore presented in Figure 4.2.

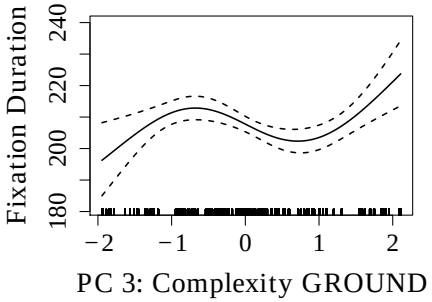

Figure 4.2: Fixation Duration results for Early Locative sentences: GROUND noun. Effect of PC3: Complexity GROUND.

As can be seen in Figure 4.2, the general trend of the effect of `PC3` is upward, with longer fixation durations for visually complex GROUND nouns as compared to visually simple GROUND nouns. However, as was the case for the complex non-linear effect of `PC6` (i.e, entropy) on the position of the initial fixation on the GROUND noun in Experiment 1, a majority of the locative phrases (70.04%) have values of `PC3` that are between -1 and 1. For these predictor values, a greater visual complexity leads to shorter fixation durations. Therefore, uncertainty remains about the nature of the effect of `PC3`. We return to this issue in our discussion of the results for `Late Locative` sentences.

4.3.3.3 Nominal

Analogous to the models for the probability of a fixation on the preposition and the GROUND noun, the `Fixation Probability` model for the topological nominal revealed a random effect of `Participant` (`Participant`: $\chi^2 = 212.085$, $p < 0.001$). Additionally, we observed an effect of `Trial` ($\chi^2 = 12.228$, $p < 0.001$), with a lower probability of a fixation on the topological nominal near the end of the experiment. This time, therefore, the effect of `Trial` is qualitatively similar to equivalent effect of `Trial` observed in Experiment 1.

The probability of a fixation on the topological nominal was furthermore influenced by lexical properties of the preceding GROUND noun, the topological nominal itself and the locative phrase as a whole. The effects of the principal components encoding these properties were highly similar to the effects reported for Experiment 1 above. First, fixation probabilities were lower for two-character (`PC8`: $\chi^2 = 4.170$, $p = 0.041$) visually complex (`PC3`: $\chi^2 = 9.770$, $p = 0.002$) GROUND nouns than for one-character visually simple GROUND nouns. Consistent with the absence of an effect of `PC8` on the probability of a fixation on the nominal in Experiment 1, the support for the effect of `PC8` is weak. Second, the probability of a fixation was greater for infrequent (`PC1`: $\chi^2 = 268.201$, $p < 0.001$) two-character (`PC11`: $\chi^2 = 14.892$, $p < 0.001$) topological nominals than for frequent single-character topological nominals. Third, the topological nominal was fixated on less often in frequent locative phrases (`PC4`: $\chi^2 = 37.788$, $p < 0.001$) as compared to infrequent locative phrases.

Furthermore, we found an effect of a predictor that did not reach significance in the equivalent model in Experiment 1: entropy (`PC6`: $\chi^2 = 4.345$, $p = 0.037$). High values of `PC6` indicate greater uncertainty about the identity of the topological nominal given the GROUND noun. In locative phrases with high entropy the topological nominal thus provides more information than in locative phrases with low entropy. Therefore, the observed positive relation between `PC6` and fixation probability is in line with the results reported above: the greater the amount of information provided by a word, the higher the probability of a fixation on that word.

When a topological nominal was fixated on, the `Fixation Position` varied considerably between participants (`Participant`: $F = 11.989$, $p < 0.001$) and items (`Items`: $F = 0.592$, $p < 0.001$). As was the case for fixations on the GROUND noun, fixations on the topological nominal were further into the word when the previous fixation was closer to the left edge of the topological nominal (`Partial Saccade`

Length: $F = 354.798$, $p < 0.001$) and when previous fixation durations were shorter (Previous Fixation Duration: $F = 5.655$, $p < 0.001$).

Furthermore, the fixation position model for the topological nominal revealed effects of three lexical properties of the topological nominal itself. As we had come to expect, fixations were further into the topological nominal when topological nominals consisted of two characters (PC11: $F = 129.906$, $p < 0.001$), were visually complex (PC7: $F = 2.471$, $p = 0.039$), or were infrequent (PC1: $F = 18.809$, $p < 0.001$). In addition, we observed an effect of phrase frequency: fixations were further into the topological nominal when the frequency of the locative phrase as a whole was low as compared to when the frequency of the locative phrase was high (PC4: $F = 14.821$, $p < 0.001$).

Lexical properties of the GROUND noun also co-determined fixation positions for the topological nominal. Consistent with the results for Experiment 1, longer GROUND nouns led to more rightward fixation positions (PC8: $F = 13.613$, $p < 0.001$). We also found an effect of the picture size of the GROUND noun (PC10: $F = 3.826$, $p = 0.025$) that was qualitatively similar to the effect of visual complexity (PC3) on the fixation position of initial fixations on the topological nominal in Experiment 1: fixation positions on the topological nominal were more leftward for GROUND nouns with greater picture sizes.

Finally, we observed an effect of the frequency of the GROUND noun in SUBTLEXCH (PC9: $F = 12.549$, $p < 0.001$). Fixations were further into the topological nominal for more frequent GROUND nouns. A possible interpretation of the positive relation between the frequency of the GROUND noun and the position of the first fixation on a topological nominal is the decreased probability of a fixation on the GROUND noun for high frequency GROUND nouns as compared to low frequency GROUND nouns. As a result, less pre-processing of the topological nominal through parafoveal preview is possible for high frequency GROUND nouns as compared to low frequency GROUND nouns. The amount of unprocessed information provided by the topological nominal thus is higher for high frequency GROUND nouns as compared to low frequency GROUND nouns, which results in more rightward fixation positions on the topological nominal.

The Fixation Duration model for initial fixations on the topological nominal revealed random effects of Participant ($F = 16.094$, $p < 0.001$) and Item ($F = 0.202$, $p = 0.0498$). The random effect of Item, however, lost significance when observations with residuals further than 2.5 standard deviations from the residual

mean were removed from the model ($F = 0.201$, $p = 0.051$). In addition, we observed effects of `Previous Fixation Duration` ($F = 3.849$, $p = 0.011$) and `Saccade Length` ($F = 9.006$, $p < 0.001$). Fixation durations were longer when the previous fixation duration was longer and the incoming saccade size was larger.

We furthermore observed effects of lexical properties of the GROUND noun and of the topological nominal itself. Fixation durations on the topological nominal were shorter for visually complex GROUND nouns as compared to visually simple GROUND nouns (`PC3`: $F = 16.080$, $p < 0.001$). Again, a possible explanation for the nature of this effect comes from the amount of pre-processing that is available through parafoveal preview for visually complex GROUND nouns as compared to visually simple GROUND nouns. Visually complex GROUND nouns were fixated on more often than visually simple GROUND nouns and therefore allow for increased parafoveal preview. Therefore, topological nominals could be pre-processed to a greater extent for visually complex GROUND nouns as compared to visually simple GROUND nouns.

Two lexical properties of the topological nominal itself influenced fixation durations. As expected, fixation durations were shorter for more frequent topological nominals as compared to less frequent topological nominals (`PC1`: $F = 20.817$, $p < 0.001$). Interestingly, fixation durations were also shorter for two-character topological nominals than for one-character topological nominals (`PC11`: $F = 16.067$, $p < 0.001$). The effect of topological nominal length thus opposite in sign compared to the effect of topological nominal length observed for fixation durations on the topological nominal in Experiment 1. Given that both effects are highly significant, the opposite nature of the topological nominal length effects for Experiment 1 and for the locative phrases in `Early Locative` sentences in Experiment 2 has to be taken seriously.

Nominals are the final word in a locative phrase. Given that locative phrases were presented in isolation in Experiment 1, locative phrases were the last word of the linguistic input in Experiment 1. By contrast, locative phrases are embedded in sentential contexts in Experiment 2. In the `Early Locative` sentences discussed here, the topological nominal is followed by a verb. Following our reasoning with respect to the facilitatory effect of the length of the GROUND noun on the duration of a fixation on the GROUND noun, a potential explanation for the facilitatory effect of the length of the topological nominal on the duration of a fixation on the topological nominal is that two-character topological nominals may have more specific

meanings than one-character topological nominals. Two-character topological nominals may thus yield stronger expectations about the upcoming verb. As a result, pre-processing of the verb through parafoveal preview may be more efficient for two-character topological nominals as compared to one-character topological nominals, which would result in shorter fixation durations. If this explanation is correct, we would expect the effect of topological nominal length for `Late Locative` sentences – in which the topological nominal is the last word – to be similar to the effect of topological nominal length in Experiment 1, with longer fixation durations for two-character nominals as compared to one-character nominals. As we demonstrate below, this prediction is borne out.

4.3.4 Results Late Locative sentences

The results for `Late Locative` sentences in Experiment 2 are presented in Table 4.4. The probabilities of a fixation on the preposition (51.90%) and the GROUND noun (87.76%) were similar to the corresponding probabilities for `Early Locative` sentences (preposition: 58.72%; GROUND noun: 92.28%). Whereas the probability of a fixation on the topological nominal was 80.86% for `Early Locative Sentences`, however, only 38.25% of all topological nominal tokens were fixated on in `Late Locative` sentences. The probability of fixation on the topological nominal for `Late Locative` sentences thus is similar to the probability of a fixation on the topological nominal for the locative phrases presented in isolation in Experiment 1 (34.51%). Given the fact that the topological nominal was the final word of the stimulus of the locative phrases in Experiment 1 and of the `Late Locative` sentences in Experiment 2 – but not for the `Early Locative` sentences in Experiment 2 – this pattern of results is as expected.

4.3.4.1 Preposition

We observed random effects of both `Participant` ($\chi^2 = 218.485$, $p < 0.001$) and `Item` ($\chi^2 = 50.464$, $p = 0.004$) on the `Fixation Probability` for the preposition. As was the case for `Early Locative` sentences, we furthermore found a facilitatory effect of `Trial` ($\chi^2 = 8.273$, $p = 0.004$), with fewer fixations on prepositions near the end of the experiment. Finally, we observed an effect of the frequency of the topological nominal (`PC1`: $\chi^2 = 5.329$, $p = 0.021$). Prepositions were fixated on more often when the topological nominal of the locative phrase was more frequent. As before, therefore, additional resources are allocated to the current word when the upcoming words provide less information.

Table 4.4: Summary of results for *Late Locative* sentences in Experiment 2. Plus symbols indicate a positive relation between predictor and response variable, minus symbols indicate a negative relation. The number of symbols corresponds to the significance of the effects, with $p < 0.001$ for three symbols, $p < 0.01$ for two symbols and $p < 0.05$ for one symbol. Round brackets indicate that the main effect of a predictor loses significance when random by-participant slopes for that predictor are added to the model.

	Preposition			Ground			Nominal		
	prob.	pos.	dur	prob.	pos.	dur.	prob.	pos.	dur.
GROUND Noun									
PC8: Length				+++	+++		+	+++	
PC3: Visual Complexity				+++	+++	+++	−−		−−
PC10: Picture Size				+	++				
PC2: Frequency				−−−					
PC9: Frequency (SUBTLEX-CH)					−		++		
Nominal									
PC11: Length					+		+++	+++	+
PC7: Visual Complexity		+					++		
PC1: Frequency	+				−−−		−−−	−−−	−−−
Construction									
PC4: Phrase Frequency				−	(−)	−	−−−		−−
PC6: Entropy									
PC5: Relative Entropy							++		

The `Fixation Position` of fixations on the preposition varied between participants (`Participant`: $F = 3.294$, $p < 0.001$) and items (`Item`: ($F = 0.228$, $p = 0.028$). Furthermore, fixations were further into the preposition when previous fixations were longer (`Previous Fixation Duration`: $F = 6.430$, $p = 0.011$) and closer to the left edge of the preposition (`Partial Saccade Length`: $F = 146.494$, $p < 0.001$). Finally, we found an effect of the visual complexity of the topological nominal on the position of fixations on the preposition (`PC7`: $F = 3.789$, $p = 0.016$). Fixations were further into the preposition when the upcoming GROUND noun was visually complex.

We did not observe lexical predictor effects on `Fixation Duration` for initial fixations on the preposition. We did, however, find a random effect of `Participant` ($F = 16.768$, $p < 0.001$), as well as effects of `Trial` ($F = 3.347$, $p = 0.021$) and `Saccade Length` ($F = 9.120$, $p < 0.001$). Fixation durations were shorter near the end of the experiment and when the size of the incoming saccade was smaller.

4.3.4.2 GROUND noun

As was the case for the probability of a fixation on the preposition, the model for the `Fixation Probability` for the GROUND noun revealed random effects for `Participant` ($\chi^2 = 436.720$, $p < 0.001$) and `Item` ($\chi^2 = 34.913$, $p = 0.015$) and an effect of `Trial` ($\chi^2 = 19.308$, $p < 0.001$). Again, the probability of a fixation decreased throughout the experiment.

The effects of lexical predictors on the probability of a fixation on the GROUND noun were highly similar for `Early Locative` and `Late Locative` phrases. As was the case for `Early Locative` phrases, the probability of a fixation was greater for two-character (`PC8`: $\chi^2 = 14.382$, $p < 0.001$) visually complex (`PC3`: $\chi^2 = 40.268$, $p < 0.001$) GROUND nouns with a bigger picture size (`PC10`: $\chi^2 = 4.346$, $p = 0.037$) as compared to visually simple single-character GROUND nouns with a smaller picture size. In addition, frequent GROUND nouns were fixated on less often than infrequent GROUND nouns (`PC2`: $\chi^2 = 29.083$, $p < 0.001$). Again, the probability of a fixation on a word thus is a function of the amount of information it provides.

The model of the fixation probability for the GROUND noun also revealed an effect of `PC4`: the frequency of the locative phrase as a whole ($\chi^2 = 4.632$, $p = 0.031$). GROUND nouns were fixated on less often in frequent locative phrases than in infrequent locative phrases. The effect of phrase frequency on the probability of a fixation on the GROUND noun is the earliest effect of phrase frequency observed

in the current study. The increased information provided by the preceding AGENT noun, the FIGURE noun and verb in `Late Locative` sentences thus helps lexical properties of the locative phrase as a whole become available at an earlier point in time.

Similar to probability of a fixation on the GROUND noun, the `Fixation Position` of the initial fixation on the GROUND noun differed between participants (`Participant`: $F = 19.272$, $p < 0.001$) and items (`Item`: ($F = 1.104$, $p < 0.001$). Consistent with the findings for `Early Locative` sentences, fixations positions were more rightward when the previous fixation was shorter (`Previous Fixation Duration`: $F = 6.743$, $p < 0.001$) and closer to the left edge of the GROUND noun (`Partial Saccade Length`: $F = 513.167$, $p < 0.001$). Furthermore, fixation positions were further into the GROUND noun near the end of the experiment than at the start of the experiment (`Trial`: $F = 8.355$, $p < 0.001$).

The effects of lexical predictors on the position of a fixation provide further support for the idea that fixation positions are driven to a considerable extent by a dynamic allocation of resources to information-rich areas. As before, the greater the amount of information provided by the GROUND noun, by the upcoming topological nominal, and by the phrase as a whole, the more rightward the position of a fixation. We found a positive relation between fixation position and the length (`PC8`: $F = 39.421$, $p < 0.001$), the visual complexity (`PC3`: $F = 7.044$, $p < 0.001$), and the picture size (`PC10`: $F = 8.743$, $p = 0.003$) of the GROUND noun, as well as between `Fixation Position` and the length of the topological nominal (`PC11`: $F = 4.203$, $p = 0.041$). By contrast, we found a negative relation between the fixation position on the GROUND noun and the frequency of the GROUND noun in SUBTLEX-CH (`PC9`: $F = 4.332$, $p = 0.037$), the frequency of the topological nominal in the SCCoW and the Gigaword corpus (`PC1`: $F = 34.732$, $p < 0.001$) and the frequency of the locative phrase as a whole (`PC4`: $F = 4.929$, $p = 0.026$). However, the statistical support for the effects of `PC11` and `PC9` was limited, whereas the effect of `PC4` lost significance when by-participant random slopes for `PC4` were added to the model ($F = 3.826$, $p = 0.051$).

The third response variable, `Fixation Duration`, showed significant random effects of `Participant` $F = 37.045$, $p < 0.001$) and `Item` ($F = 0.465$, $p < 0.001$). In addition, fixation durations were shorter near the end of the experiment (`Trial`: $F = 6.749$, $p = 0.001$) and for smaller incoming saccade sizes (`Saccade Length`: $F = 5.859$, $p < 0.001$), as well as for fixations that were further into the word (`X`

Position: $F = 3.423$, $p = 0.016$) and lower on the screen (Y Position: $F = 2.941$, $p = 0.0496$). However, the support for the effect of Y Position was weak.

We observed effects of two lexical predictors on the duration of initial fixations on the GROUND noun. First, visually complex GROUND nouns were fixated on longer than visually simple GROUND nouns (PC3: $F = 13.389$, $p < 0.001$). This sheds further light on the complex non-linear effect of PC3 on the duration of fixations on the GROUND nouns in Early Locative sentences. While the overall trend of the effect of PC3 was upward, we found an opposite pattern of results for the middle of the PC3 range. The reversal of the effect of medium predictor values is not replicated for Late Locative sentences. We therefore conclude that there is a positive relation between PC3 and Fixation Duration for initial fixations on the GROUND noun when reading locative phrases embedded in sentences. Hence, a greater amount of information leads to longer fixation durations.

The amount of information in the locative phrase as a whole also influences fixation durations on the GROUND noun, as indicated by an effect of phrase frequency (PC4: $F = 5.526$, $p = 0.019$). Surprisal is greater for low frequency phrases as compared to high frequency phrases. The fact that fixation durations on the GROUND noun were shorter for high frequency locative phrases as compared to low frequency locative phrases thus is consistent with our expectations.

4.3.4.3 Nominal

The Fixation Probability for the topological nominal differed significantly between participants (Participant: $\chi^2 = 530.564$, $p < 0.001$) and items (Item: $\chi^2 = 52.150$, $p = 0.002$). Furthermore, we observed effects of principal components that encode lexical properties of the GROUND noun, the topological nominal itself, and the locative phrase as a whole. The probability of a fixation on the topological nominal was greater when the preceding GROUND noun had a higher frequency in SUBTLEX-CH and, as a result, was fixated on less often (PC9: $\chi^2 = 6.856$, $p = 0.009$).

The results for the visual complexity of the GROUND noun were mixed. On the one hand, a greater visual complexity of the GROUND noun led to fewer fixations on the topological nominal (PC3: $\chi^2 = 10.142$, $p = 0.001$). This finding is consistent with the result of visual complexity on the probability of a fixation on the topological nominal for Early Locative sentences and suggests that a higher probability of a fixation on the GROUND noun leads to a lower fixation probability for the topological nominal. On the other hand, however, topological nominals were fixated on more

frequently when the preceding GROUND noun was a two-character word than when it was a one-character word (PC8: $\chi^2 = 5.884$, $p = 0.015$). The effect of PC8 thus is opposite to the effect of PC8 for Early Locative sentences. However, given the relatively subtle nature of the effect of PC8 and the absence of an effect of PC8 on the probability of a fixation on the nominal for the locative phrases presented in isolation in Experiment 1, this reversal is not statistically robust.

The results for lexical properties of the topological nominal are in line with our expectations. Two-character topological nominals were fixated on more often than one-character topological nominals (PC11: $\chi^2 = 88.589$, $p < 0.001$) and fixation probabilities were higher for visually complex topological nominals than for visually simple topological nominals (PC7: $\chi^2 = 7.893$, $p = 0.005$). Furthermore, we observed an increased fixation probability for low frequency topological nominals as compared to high frequency topological nominals (PC1: $\chi^2 = 593.604$, $p < 0.001$). As before, these results indicate that the probability of a fixation on a word is proportional to the surprisal associated with that word and thus to the amount of information it provides.

Finally, we found two construction-level effects on the probability of a fixation on the topological nominal. As was the case for Early Locative sentences, the topological nominal was fixated on less frequently in high frequency locative phrases as compared to low frequency locative phrases (PC4: $\chi^2 = 78.369$, $p < 0.001$). In addition, we observed a relative entropy effect (PC5: $\chi^2 = 7.805$, $p = 0.005$), with a higher fixation probability for GROUND nouns with a high relative entropy than for GROUND nouns with a low relative entropy. The amount of information provided by the topological nominal is greater for GROUND nouns with a high relative entropy as compared to GROUND nouns with a low relative entropy. Hence, the positive relation between fixation probability and relative entropy is exactly as expected.

The Fixation Position model for the topological nominal revealed random effects of Participant $F = 4.384$, $p < 0.001$) and Item ($F = 0.542$, $p < 0.001$). Fixations were further into the topological nominal if the preceding fixation was closer to the topological nominal (Partial Saccade Length: $F = 148.313$, $p < 0.001$) and if a fixation was lower on the screen (Y Position: $F = 3.296$, $p = 0.023$). Furthermore, fixations were less far into the word at the start of the experiment than at the end of the experiment (Trial: $F = 3.436$, $p = 0.015$).

The effects of the lexical predictors on the position of the first fixation on the topological nominal were in line with the results for Experiment 1 and for the `Early Locative` sentences reported above. Fixations were further into the topological nominal if the topological nominal itself (`PC11`: $F = 47.388$, $p < 0.001$) or the preceding GROUND noun (`PC8`: $F = 15.066$, $p < 0.001$) was longer and less far into the topological nominal if the topological nominal was more frequent (`PC1`: $F = 41.544$, $p < 0.001$).

The `Fixation Duration` model for the topological nominal revealed that fixation durations differed significantly between participants (`Participant`: $F = 8.774$, $p < 0.001$). We furthermore found effects of 5 control variables. As was the case for initial fixations on the preposition and the GROUND noun, fixations durations for initial fixations on the topological nominal decreased throughout the experiment (`Trial`: $F = 15.543$, $p < 0.001$). Fixation durations were longer for fixations that were less far into the word (`X Position`: $F = 44.565$, $p < 0.001$) and near the vertical center of the word (`Y Position`: $F = 3.336$, $p = 0.032$). Additionally, fixation durations were shorter when previous fixations were shorter (`Previous Fixation Duration`: $F = 5.110$, $p = 0.018$). Interestingly, fixation durations were shorter for larger incoming saccade sizes as well (`Saccade Length`: $F = 2.998$, $p = 0.026$).

The effect of incoming saccade length is opposite to the effects of saccade length on fixation durations observed thus far. Typically, we found that incoming saccade sizes lead to shorter fixations, presumably due to increased pre-processing through parafoveal preview. By contrast, smaller incoming saccade sizes result in longer fixations on topological nominals in `Late Locative` sentences. The reversal of the effect of `Saccade Length` may be related to the additional information provided by the preceding AGENT noun, FIGURE noun, and verb in `Late Locative` sentences. This information may allow for fully optimized fixation planning based on the expected amount of information in the topological nominal. A more leftward position of the previous fixation, then, may indicate that the topological nominal provides less information and can be processed more quickly.

Fixation durations were furthermore co-determined by the visual complexity of the GROUND noun (`PC3`: $F = 7.712$, $p = 0.006$), with shorter fixation durations on the topological nominal for visually complex GROUND nouns as compared to visually simple GROUND nouns. This effect is qualitatively similar to the effect of `PC3` on the duration of fixations on the topological nominal in `Early Locative` sentences and is likely to reflect increased parafoveal preview due to the higher probability of

a fixation on the GROUND noun and the more rightward fixation position and longer fixation duration of such a fixation for visually complex GROUND nouns as compared to visually simple GROUND nouns.

Lexical properties of the topological nominal itself also had an effect on the duration of the initial fixation on the topological nominal. Again, we found a positive relation between the amount of information provided by the topological nominal and fixation duration, with longer fixation durations for infrequent (`PC1`: $F = 66.984$, $p < 0.001$) two-character (`PC11`: $F = 5.459$, $p = 0.020$) topological nominals as compared to frequent topological nominals that consist of a single character. This effect is in line with our prediction that the effect of topological nominal length should be opposite for locative phrases embedded in `Early Locative` and in `Late Locative` sentences (see Section 4.3.3.3). Finally, we observed an effect of the frequency of the locative phrase as a whole (`PC4`: $F = 8.113$, $p = 0.004$), with shorter fixation durations on the topological nominal for more frequent locative phrases.

4.3.5 Discussion

Whereas locative phrases were presented in isolation in Experiment 1, we embedded locative phrases in sentence contexts in Experiment 2. We investigated two types of sentence structures: `Early Locative` sentences and `Late Locatives` sentences. As was the case for Experiment 1, we analyzed the fixation probability, fixation position and fixation duration for the preposition, the GROUND noun and the topological nominal using GAMMs.

For the locative phrases presented in isolation in Experiment 1 pre-processing through parafoveal preview was ubiquitous. Lexical properties of the topological nominal and the locative phrase as a whole had a large influence on the fixation patterns on the GROUND noun. For the `Early Locative` sentences in Experiment 2, parafoveal preview effects were much less prominent. We observed subtle effects of the visual complexity of the topological nominal and relative entropy on the position of a fixation of the GROUND noun only. For `Late Locative` sentences the influence of lexical properties of the topological nominal and the phrase as a whole on fixation patterns for the GROUND noun was somewhat greater due to the increased information provided by the preceding AGENT noun, FIGURE noun and verb. Nonetheless, as a whole, pre-processing through parafoveal preview was less prominent for locative phrases embedded in both `Early Locative` and `Late Locative` sentences than for locative phrases presented in isolation.

There are at least three possible explanations for the decreased role of parafoveal preview in Experiment 2. First, due to a change in the setup of the experimental lab the experiment was presented on a 25.9 inch flatscreen monitor rather than on the 17 inch LCD monitor used for Experiment 1. Consequently, the number of characters per degree of visual angle was somewhat lower in Experiment 2 (0.66) than in Experiment 1 (0.75), despite the use of a smaller font in Experiment 2 (28 point) than in Experiment 1 (36 point). Therefore, the amount of information available through parafoveal preview was 14% lower in Experiment 2 than in Experiment 1.

Second, the experimental task was substantially harder in Experiment 2 than in Experiment 1. In both experiments participants were asked to judge the correctness of a translation of the experimental item to English. However, in Experiment 1 incorrect translations differed from the correct translation with respect to the GROUND noun or the topological nominal only. By contrast, incorrect translations in Experiment 2 differed with respect to either the GROUND noun, the topological nominal, the FIGURE noun, or the verb. The experimental task in Experiment 2 required participants to pay close attention to individual words. This may have resulted in increased attention to information in the fovea at the cost of pre-processing based on information in the parafovea.

Third, the reduction in parafoveal processing could be intrinsic to reading locative phrases in sentence contexts rather than in isolation. Under this hypothesis, the limited ecological validity of the experimental design for Experiment 1 may have resulted in an artificial increase of pre-processing through parafoveal preview as compared to the amount of pre-processing that is typical in everyday language processing.

The current data do not allow us to distinguish between these explanations for the quantitative difference in parafoveal preview effects between Experiment 1 and Experiment 2. It could well be the case that all three explanations contribute to this difference to some extent. Further research is necessary to gain more insight into the exact extent to which parafoveal preview influences eye movements during locative phrase reading. The question that remains open, however, is one of magnitude, rather than actuality. Even when locative phrases were embedded in sentences we found ample evidence for the idea that pre-processing through parafoveal preview plays an important role in the processing mechanisms that underlie locative phrase reading in Mandarin Chinese.

The second base mechanism identified in Experiment 1 was a dynamic allocation of resources to information-rich areas. The results for both `Early Locative` sentences and `Late Locative` sentences in Experiment 2 confirmed that the search for information plays a pervasive role in fixation planning in Mandarin Chinese. The amount of information provided by the current word, by the preceding word, and by the upcoming word co-determines the fixation probability, the fixation position, and the fixation duration for the current word to a considerable extent.

Experiment 1 revealed effects of phrase frequency and entropy on the processing of locative phrases. Experiment 2 provided further support for the effects of phrase frequency and entropy on the fixation patterns during locative phrase reading. Phrase frequency effects were limited to fixation patterns on the topological nominal for `Early Locative` phrases, but were present for fixation patterns on both the GROUND noun and the topological nominal for `Late Locative` phrases. The additional information provided by the preceding AGENT noun, FIGURE noun, and verb in `Late Locative` sentences thus leads to an earlier availability of phrase-level lexical information. We furthermore observed an effect of entropy on the probability of a fixation on the topological nominal in `Late Locative` sentences.

In addition to the effects of phrase frequency and entropy, we observed effects of a third predictor at the construction level: relative entropy. Relative entropy describes the prototypicality of the frequency distribution of topological nominals for a GROUND noun. When the topological nominal paradigm of the preceding GROUND noun had an atypical frequency distribution, the fixation position on the GROUND noun was more rightward for `Early Locative` sentences, whereas the probability of a fixation on the topological nominal was greater for `Late Locative` sentences. Greater uncertainty about the identity of the topological nominal thus leads to the use of additional resources. Previously, Baayen et al. (2011) and Hendrix et al. (2016) observed relative entropy effects for prepositional phrases in English. The current effect of relative entropy is, to our knowledge, the first relative entropy effect documented for Mandarin Chinese.

4.4 Quantitative analysis: gradient boosting machines

The GAMM analyses provide detailed insights into the qualitative nature of the effects of the principal components and the control variables on the eye fixation patterns during locative phrase reading. However, they provide less conclusive evidence about the quantitative contribution of predictors. In this section we complement the findings reported above with an investigation of the explanatory power of the control variables and the principal components for the eye fixation patterns through an analysis using gradient boosting machines (GBMs).

For each of the GAMMs reported above, we fitted an analogous GBM using version $0.4-3$ of the `xgboost` package for the statistical software R (T. Chen et al., 2015). The predictors entered into each GBM model were identical to those entered into the corresponding GAMM. All GBMs consisted of 100 trees and were fitted with the standard parameter settings. For the `Fixation Duration` and `Fixation Position` models we used linear regression as the objective function, whereas for the `Fixation Probability` GBMs we used logistic regression as the objective function. Outlier removal for the response variable was identical to the outlier removal for the response variables for the corresponding GAMMs. No predictor outliers were removed prior to analysis.

The `xgboost` packages reports variable importances as relative influences (i.e., percentage of the total information gain for all variables in the model), which sum up to 100% for each model. This leads to equally high relative influences of predictors for GBMs that explain a lot of variance and for GBMs that explain little variance. To obtain insight into the absolute contribution of a predictor, we calculated a second variable importance measure by multiplying relative influences with the proportion of variance explained by a GBM model under 10-fold cross validation. We refer to this second measure of variable importance as "absolute influence". Below, we discuss the overall pattern of results that emerged from the GBM analysis. Exact relative influences and absolute influences for each GBM are reported in Appendix B. Predictors with an influence of zero are omitted from these model summaries.

Table 4.5 presents the percentage of variance explained for each of the GBM models fitted to the eye fixation patterns for the locative phrases presented in isolation in Experiment 1 and for the locative phrases in the `Early Locative` and `Late Locative` sentences in Experiment 2. The average percentage of variance explained by

Table 4.5: Percentage of variance explained under 10-fold cross-validation by the GBM models for the eye fixation patterns in Experiment 1 (Isolation) and in the `Early Locative` and `Late Locative` sentences in Experiment 2.

	Preposition			GROUND Noun			Topological Nominal		
	prob.	pos.	dur	prob.	pos.	dur.	prob.	pos.	dur.
Isolation	11.93	14.42	24.90	20.80	58.07	35.11	35.28	26.58	29.64
Early Loc.	12.44	14.83	25.59	19.81	33.15	19.85	15.02	39.99	18.10
Late Loc.	6.53	21.57	23.49	21.04	39.55	25.64	37.63	34.77	19.45

the GBMs fitted to the eye fixation patterns in Experiment 1 (28.52%) was somewhat higher than the average percentage of variance explained by the GBMs fitted to the eye fixation patterns for the locative phrases in the `Early Locative` (22.09%) and `Late Locative` (25.52%) sentences in Experiment 2.

Across both experiments, the percentage of variance explained by the principal components and the control variables was greatest for the GBMs fitted to the eye fixation data for the GROUND noun (30.33%), followed by the topological nominal (28.50%) and the preposition (17.30%). The preposition is identical for all locative phrases. Predictor effects on the preposition thus are limited to parafoveal preview effects of lexical properties of the upcoming GROUND noun and topological nominal. Hence, it is expected that relatively little variance can be explained by the GBMs fitted to the eye fixation patterns on the preposition.

The variable importances of the principal components in the GBMs fitted to the eye fixation data for the preposition are shown in Figure 4.3. Depicted is the relative contribution of each principal component and each group of principal components (i.e., principal components describing lexical properties of the GROUND noun, the topological nominal and the locative construction as a whole), averaged across the GBMs for the three response variables (i.e, `Fixation Probability`, `Fixation Position`, and `Fixation Duration`) and the three experimental conditions (i.e., Experiment 1, `Early Locative` sentences in Experiment 2, and `Late Locative` sentences in Experiment 2). Control variables are omitted from Figure 4.3.

The summed average absolute influence of the principal components across the nine GBMs fitted to the eye movement data for fixations on the preposition is 1.58. As can be seen in Figure 4.3, lexical properties of the GROUND noun are most predictive for the fixation patterns on the preposition (summed average absolute influence: 0.75). The strongest individual predictor is `PC8` (i.e., the length of the GROUND

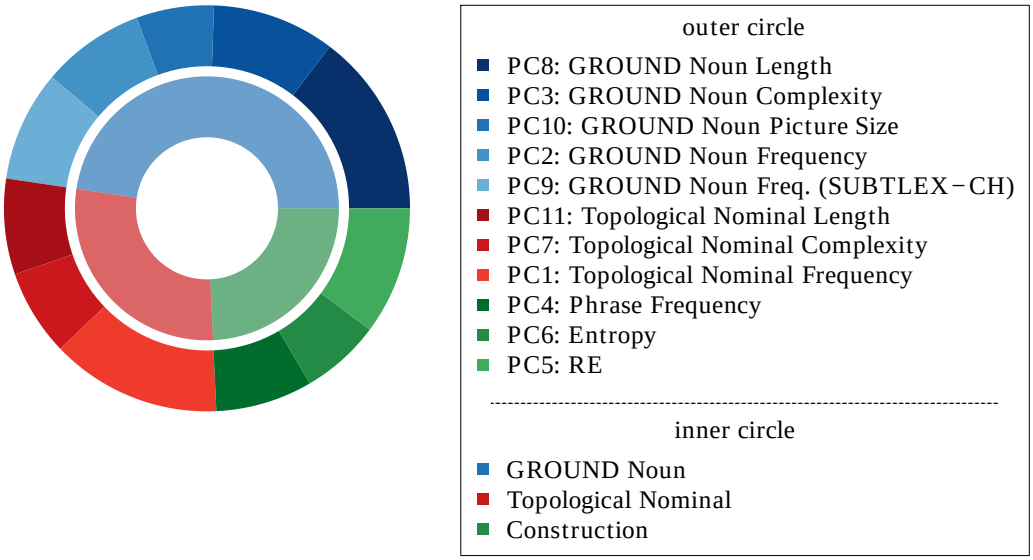

Figure 4.3: Variable importances in the GBM models fitted to the eye fixation data for initial fixations on the preposition.

noun), which has an average absolute influence of 0.23. The visual complexity of the GROUND noun (PC3; average absolute influence: 0.15), as well as the frequency of the GROUND noun in SUBTLEX-CH (PC9; average absolute influence: 0.14) and in the SCCOW (PC2; average absolute influence: 0.13) also have a substantial contribution to the explanatory power of the GBMs for the fixation patterns on the preposition.

The principal components describing lexical properties of the topological nominal (summed average absolute influence: 0.44) and the construction as a whole (summed average absolute influence: 0.38) have a more modest influence on fixation patterns on the preposition. Consistent with the results of the GAMMs, the predictor with the greatest average absolute influence that does not describe a lexical property of the GROUND noun is PC1 (i.e., the frequency of the topological nominal; average absolute influence: 0.21).

The contribution of the principal components for fixation patterns on the preposition was relatively limited (summed average absolute influence: 1.58). By contrast, the explanatory power of the principal components for fixation patterns on the GROUND noun is substantial (3.19). The relative influences of the (groups of) principal components in the GBMs fitted to the fixation data for the GROUND noun are shown in Figure 4.4. Lexical properties of the GROUND noun itself are most predictive for fixation patterns on the GROUND noun (summed average absolute

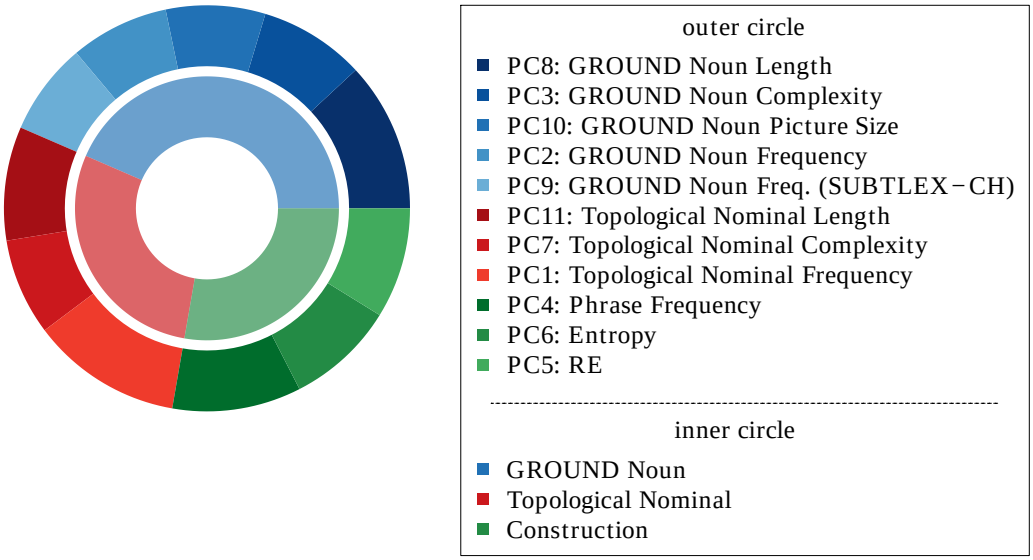

Figure 4.4: Variable importances in the GBM models fitted to the eye fixation data for initial fixations on the GROUND noun.

influence: 1.39). However, principal components encoding lexical properties of the topological nominal (summed average absolute influence: 0.92) and the construction as a whole (summed average absolute influence: 0.89) were predictive for fixation patterns on the GROUND noun as well.

As can be seen in Figure 4.4, the measure related to the GROUND noun that is most predictive for the fixation data for the GROUND noun is `PC8` (i.e., the length of the GROUND noun; average absolute influence: 0.38). The frequency of the topological nominal has a substantial contribution to the explanatory power of the GBMs as well (`PC1`; average absolute influence: 0.38), as does the length of the topological nominal (`PC11`; average absolute influence: 0.29). The top 5 of the predictors with the greatest absolute influence is completed by two measures of the construction as a whole: `PC4` (i.e., the frequency of the locative phrase as a whole; average absolute influence: 0.33) and `PC5` (i.e., the relative entropy of the GROUND noun; average absolute influence: 0.28).

Figure 4.5 presents the relative variable importances of the principal components for the fixation patterns for the topological nominal. The summed average absolute influence for the principal components is 1.58. Whereas lexical properties of the GROUND noun had the greatest summed average absolute influence in the GBMs fitted to the fixation data for the preposition and the GROUND noun, the fixation patterns on the topological nominal are influenced most by lexical proper-

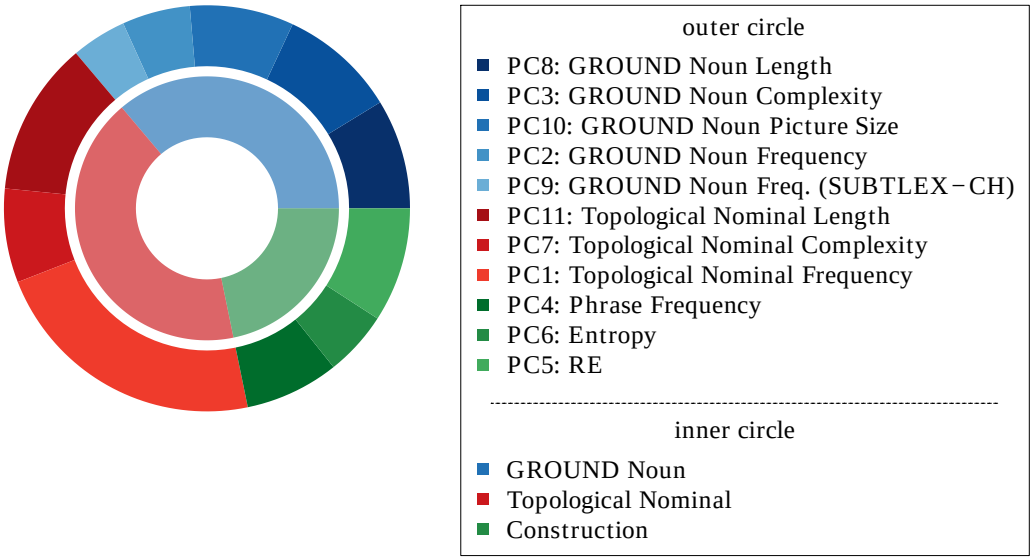

Figure 4.5: Variable importances in the GBM models fitted to the eye fixation data for initial fixations on the topological nominal.

ties of the topological nominal (summed average absolute influence: 1.21). Lexical properties of the GROUND noun (summed average absolute influence: 1.04) and the locative construction as a whole (summed average absolute influence: 0.63), however, contribute to the predictive power of the GBMs for the fixation patterns on the topological nominal as well.

The GBMs for the fixation patterns on the topological nominal are dominated by the principal components that encode the frequency (`PC1`; average absolute influence: 0.64) and the length (`PC11`; average absolute influence: 0.35) of the topological nominal itself. However, the visual complexity (`PC3`; average absolute influence: 0.27), the length (`PC8`; average absolute influence: 0.25), and the picture size (`PC10`; average absolute influence: 0.24) of the GROUND noun had explanatory power for eye fixation data for the topological nominal as well, as did the frequency of the locative phrase as a whole (`PC4`; average absolute influence: 0.22) and the relative entropy of the GROUND noun (`PC5`; average absolute influence: 0.26).

This concludes our discussion of the results from the GBM analyses of the eye fixation data. Unsurprisingly, the GBM analyses indicated eye movement patterns for the GROUND noun were mostly influenced by lexical properties of the GROUND noun, whereas eye movement patterns for the topological nominal were primarily influenced by lexical properties of the topological nominal. However, lexical proper-

ties of upcoming words co-determined fixation patterns for the preposition and the GROUND noun as well. Hence, the GBM analysis provides further support for the idea that upcoming words are pre-processed through parafoveal preview.

Consistent with the GAMM analyses reported above, the GBM analyses furthermore showed considerable contributions of phrase frequency and relative entropy to the GBMs fitted to the fixation data for the GROUND noun and the topological nominal. These findings provide additional evidence for the influence of construction-level lexical predictors on eye movement patterns during locative phrase reading that we reported in the GAMM analyses above. Overall, the results of the GBM analyses and the GAMM analyses thus converge.

4.5 General discussion

We presented the results of two phrase reading experiments in which we investigated lexical processing in Mandarin Chinese at the word level and the phrase level. In both experiments we recorded eye movement patterns while participants read locative phrases. Locative phrases are similar to prepositional phrases in English and consist of a (semantically empty) preposition, a GROUND noun and a topological nominal. In Experiment 1 locative phrases were presented in isolation. In Experiment 2 locative phrases were embedded in two types of sentential contexts that differed with respect to the position of the locative phrase in the sentence (i.e., early or late). We analyzed the probability of a fixation, the position of the initial fixation, and the duration of the initial fixation for each of the three words in a locative phrase using generalized-additive mixed-effect models (GAMMs).

The analysis of the eye fixation data resulted in three main findings. First, consistent with previous findings (Inhoff & Liu, 1997; Inhoff, 1999; J. Yang et al., 2009) and the results reported for word naming in Chapter 3, parafoveal preview provides important guidance for readers of Mandarin Chinese during locative phrase reading. Parafoveal preview refers to the process by which readers use information in the periphery of their visual field to gather information about upcoming words. In the current experiments, participants used parafoveal preview to pre-process later parts of the locative phrase and to optimize eye fixation patterns. Parafoveal preview influenced the probability of a fixation on a word in the locative phrase, as well as the position and duration of such a fixation. Predictor effects related to parafoveal preview were most prominent for the locative phrases presented in iso-

lation in Experiment 1, but were also present in Experiment 2, where the locative phrases were embedded in sentential contexts. Reading in Mandarin Chinese thus is highly anticipatory in nature.

The second main finding of the current study is that the search for information plays a pivotal role during reading. Consistent with the ideas of information theory and the findings for word naming in Chapter 3, a driving force behind fixation patterns and lexical predictor effects is a dynamic allocation of attention to information-rich areas. One-character, visually simple words with a high frequency provide less information than two-character visually complex words with a low frequency. Hence, these words are fixated on less often and less long. In addition, fixation positions are further into frequent visually simple one-character words as compared to infrequent visually complex two-character words. The influence of the continuous search for information, however, is not limited to eye fixation patterns on the current word. The amount of information provided by the upcoming words likewise co-determines fixation patterns on the current word. Typically, less resources are allocated to the current word if the upcoming words provides more information.

Third, in addition to the expected effects of visual complexity and word frequency (see e.g., Inhoff & Liu, 1998; H. M. Yang & McConkie, 1999; G. Yan et al., 2006), we observed effects of predictors related to the combinatorial properties of words. We observed robust and consistent phrase frequency effects for locative phrases presented in isolation as well as for locative phrases embedded in both types of sentential contexts. Both the GROUND noun and the topological nominal were fixated on less often in more frequent locative phrases. Additionally, fixation durations were shorter and fixations were less far into the word when the GROUND noun or the topological nominal was fixated on in a more frequent phrase.

Interestingly, 140 of the locative phrases used in the current study were not present in the 466 million word SCCoW. Similarly, 128 of the locative phrases used here did not occur in the Gigaword corpus, which consists of 718 million word tokens. The size of these corpora approaches or exceeds reasonably estimates of the upper limit of the number of words participants could have experienced in their lifetime. In order to experience 500 million words by the age of 25 (i.e., the average age of the participants in this study), participants would need to have heard or read over $3,400$ words per hour during the 16 hours of the day they were not asleep. Therefore, it is safe to assume that any given participant never experienced a substantial number of locative phrases used in the experiment. Hence, the phrase frequency

effects observed here may reflect not only linguistic experience, but also conceptual knowledge about the configurations in which FIGURE nouns and GROUND nouns occur in the world.

Furthermore, we found effects of both entropy and relative entropy. The effects of entropy and relative entropy were less prominent than the effects of phrase frequency in the GAMM analyses. Nonetheless, additional resources were allocated to the GROUND noun and the topological nominal for locative phrases with less predictable topological nominals (i.e, high entropy phrases) and for locative phrases in which the frequency distribution of the topological nominal paradigm for the GROUND noun was atypical (i.e., high relative entropy GROUND nouns). In addition, relative entropy – and to a lesser degree – entropy had substantial contributions to the explanatory power of GBMs fit to the fixation patterns for both experiments.

The current findings are consistent with Rayner et al. (2005) and H. C. Wang et al. (2010), who found effects of subjective predictability and transitional probability on eye movement patterns. To our knowledge, the effects of phrase frequency, entropy, and relative entropy are the first n-gram frequency, entropy and relative entropy effects reported for multi-word reading in Mandarin Chinese. The effects of these predictors indicate that the combinatorial properties of linguistic elements play an important role in lexical processing not only below the word level (see Chapter 3), but also at the word level. An understanding of the information-theoretic and distributional properties of the language thus is crucial for understanding language processing in Mandarin Chinese.

The current chapter is by no means an attempt at an exhaustive investigation of lexical processing above the word level. Rather, it is first exploration of phrase-level processing using a specific linguistic construction. Much more research – using a wide range of experimental paradigms and linguistic stimuli – is necessary to gain a deeper understanding of how different distributional properties influence lexical processing at and above the phrase level. Nonetheless, the results presented here provide interesting insights to the general mechanisms that underlie language processing at the phrase level in Mandarin Chinese, as reflected in the eye movement patterns for locative phrases.

5

Conclusions

Over the last decades, numerous lexical databases have been developed for well-studied languages, such as English (Coltheart, 1981), German (Heister et al., 2011) and French (New et al., 2007). Lexical databases are a useful resource for psycholinguistic researchers. The use of predictors from a lexical database allows for a more direct comparison of the results of different experiments. Furthermore, the availability of lexical databases enables researchers to efficiently control for the effects of predictors that are extrinsic to the experimental question, but that may nonetheless influence behavioural measures of language processing.

However, simplified Chinese lexical resources are scarce. Most notably, Y. Liu et al. (2007) provide word naming latencies and 15 lexical variables for nearly $2,500$ one-character words. This thesis introduced a new, large-scale lexical database for simplified Chinese: the Chinese Lexical Database (CLD). The CLD is an order of magnitude larger than existing lexical resources for simplified Chinese. It comprises 141 numerical and 23 categorical variables for $4,710$ one-character words and $25,935$ two-character words, for a total of $(30,645 * 164 =)$ $5,025,780$ data points. The lexical variables in the CLD describe orthographic, phonological and information-theoretic properties of simplified Chinese at the word level, at the character level and below the character level (e.g., for the phonetic or semantic radical). The CLD is publicly available and can be downloaded and searched at `http://www.chineselexicaldatabase.com`.[1]

[1]The website `http://www.chineselexicaldatabase.com` is password-protected until this dissertation is published. The password is `75090246`.

Chapter 2 of this dissertation provides a detailed introduction to the CLD and the lexical variables it comprises. In this chapter, I discussed the 23 categorical predictors in the CLD. These categorical predictors contain information about the type and structure of the characters in a word, about the pronunciation of the word and its characters, and about the semantic and phonetic radicals in a word's characters. Next, I presented the 141 numerical predictors in the CLD on the basis of the results of a clustering technique applied to a self-organizing map (SOM) trained on the (squared) correlation matrix for the numerical predictors. SOMs are neural networks that are trained in an unsupervised manner and that have a spatial organization. Neurons that are closer together in a SOM encode similar information. In the context of this dissertation, SOMs allowed for a low-dimensional representation of the correlational structure between the 141 numerical predictors in the CLD, in which similar predictors were topographically grouped together.

The hierarchical clustering technique applied to the CLD yielded 21 clusters that showed a remarkable degree of conceptual and distributional coherence and that allowed for a structured discussion of the numerical predictors in the CLD. For this discussion, I allocated the 21 clusters to 6 groups of clusters. Each group of clusters comprised conceptually similar predictors. The first group of clusters contained frequency measures, as well as information-theoretic measures. Frequency measures included frequency per million, contextual diversity, family size and family frequency for words, characters and phonetic radicals. Information-theoretic measures encoded information about the entropy and mutual information of characters and words, respectively. The second group of clusters encoded the visual complexity of words, characters and radicals. Visual complexity measures in the CLD include stroke counts, high-level and low-level components counts and pixel counts, as well as picture sizes. Furthermore, the lexical predictors in the second group of clusters included orthographic neighbourhood measures at different grain sizes.

The CLD furthermore contains information about the phonological properties of a word. The measures in the third group of clusters described the phonological complexity, the phonological frequency and the phonological neighbourhood density of a word and its characters. The lexical predictors in the fourth and fifth group of clusters described the consistency of the mapping from orthography to phonology and of the mapping from phonology to orthography. Variables in these groups of clusters included type counts, token counts and frequencies of homographs and homophones, both at the character-level and at the level of the phonetic radical. The

final group of clusters comprised a single cluster that contains various other lexical predictors, including measures of the frequency and complexity of the semantic radical and measures of the frequency of a word and its characters in traditional Chinese.

Chapter 3 and Chapter 4 of this thesis explored the explanatory power of the lexical predictors in the CLD for behavioural measures of language processing obtained through a word naming task and a phrase reading task. The data presented in these chapters were analyzed using state-of-the-art statistical techniques. I established the quantitative contribution of lexical predictors through the use of gradient boosting machines (GBMs; J. H. Friedman, 2001, 2002). GBMs are a tree-based machine learning technique that is similar in spirit to random forests (Strobl et al., 2009). Whereas trees in random forests are grown independently, however, trees in GBMs are fitted sequentially – with each tree being fitted to the residuals of the ensemble of previous trees.

To gain further insight into the qualitative nature of predictor effects I used generalized additive mixed-effect models (GAMMs; Hastie & Tibshirani, 1986; S. Wood, 2006; S. N. Wood, 2011). GAMMs are regression models that allow for non-linear main effects and interactions without any predefined structure, while controlling for random effects of, for instance, participant and item. Where necessary, I used principal component analyses to overcome the problem of collinearity by conflating highly correlated lexical variables into a single predictor. The use of these statistical techniques allowed for an in-depth investigation of the effects of the categorical and numerical predictors in the CLD on the behavioural measures obtained during a word naming task and a phrase reading task.

Chapter 3 presents the results of a single-participant study in which a native reader of simplified Chinese read all the 30,645 words in the CLD. I analyzed three measures of lexical processing collected during this experiment: naming latencies, pronunciation durations and eye fixation durations. The analyses for these three measures provided detailed insights into the effects of a wide range of predictors in the CLD. First, consistent with previous findings (c.f., Y. Liu et al., 2007; Sze et al., 2014; G. Yan et al., 2006), I found strong effects of the frequency of a word and its characters on naming latencies and eye fixation durations, as well as a word frequency effects on pronunciation durations. Furthermore, I observed effects of the frequency of the semantic radical on the eye fixation patterns.

The analysis of multiple behavioural correlates of language processing allows for insights into the nature of lexical processing that are not available through the analysis of a single dependent variable. For example, the character and word frequency effects for the naming latencies are in principle consistent with parallel dual-route models, which predict that a two-character word and its characters are activated simultaneously (see, e.g., Schreuder & Baayen, 1995; Baayen & Schreuder, 1999, 2000). However, the analysis of the eye fixation durations indicated that character frequency effects temporally precede word frequency effects. This finding cannot straightforwardly be explained by parallel dual-route models. The analysis of the eye fixation durations thus provided information about the nature of lexical processing that was not available through an analysis of the naming latencies.

Second, I observed effects of the visual complexity of a word and its characters on all three measures of lexical processing (c.f., Y. Liu et al., 2007; Lee et al., 2015). The grain size at which the effects of visual complexity were best measured differed between response variables. Pixel counts were most predictive for eye fixation durations, stroke counts had most explanatory power for naming latencies and pronunciation durations were co-determined to the greatest extent by the picture size of the linguistic unit. Third, phonological properties of words and their characters influenced pronunciation durations, which showed robust effects of phoneme and diphone frequency measures, phonological complexity measures, and phonological neighbourhood density measures. In addition, the number of words in which a phonetic radical was pronounced the same had a strong influence on the naming latencies for one-character words (c.f., Y. Liu et al., 2007; Seidenberg, 1985; Hue, 1992).

Frequency, visual complexity and – to a lesser degree – phonological predictors are well-studied in psycholinguistic literature for Chinese. The influence of information-theoretic measures on behavioural measures of lexical processing has received much less attention (c.f., H. C. Wang et al., 2010). Here, I observed robust effects of the entropy of the first and second character (i.e., the entropy over the frequency of two-character words with a specific first or second character) on the naming latencies, as well as on the eye fixation durations. Interestingly, naming latencies and fixation durations were shorter when words contained characters with a high entropy, presumably due to the increased orthography-to-phonology consistency of characters that combine with many other characters.

Furthermore, I observed effects of relative entropy (i.e., the prototypicality of the frequency distribution of two-characters words with a specific first or second character) and of the entropy over the character frequencies on both the naming latencies and the eye fixation durations. Naming latencies and eye fixation durations were shorter for high values of relative entropy and longer for high values of the entropy over the character frequencies. To my knowledge, the entropy effects observed here are the first effects of entropy reported in the psycholinguistic literature of simplified Chinese. These effects illustrate that a comprehensive understanding of lexical processing in Chinese requires a thorough understanding of the information-theoretic properties of the language.

Chapter 3 provided further evidence for the validity of the results of a single-participant study (c.f., Pham & Baayen, 2015). The results for the naming latencies and pronunciation durations of a second participant (i.e., the author of this dissertation) showed a remarkable degree of convergence with the effects discussed above. Hence, the key effects reported in Chapter 3 are likely to generalize to other highly educated native readers of simplified Chinese. Chapter 3 thus provides detailed insights into lexical processing in Chinese at and below the word level.

Chapter 3 presented the results of an investigation of lexical processing at and *below* the word level in a single-participant study. In Chapter 4 I investigated lexical processing at and *above* the word level in a multiple-participant study. The construction under investigation is the Chinese equivalent of prepositional phrases in English: the locative phrase. Locative phrases consist of three elements: the semantically empty preposition 在, a GROUND noun and a topological nominal. The locative phrases were presented to participants in two experiments. In Experiment 1 locative phrases were presented in isolation, whereas in Experiment 2 locative phrases were embedded in sentences, either in early position or in late position. I recorded and extracted eye movement patterns during both experiments and analyzed the probability of a fixation, as well as the position and the duration of the initial fixation for each of the three words in a locative phrase.

As expected, the results for the locative phrase reading experiments revealed effects of the frequency and the visual complexity of both the GROUND noun and the nominal (c.f., Inhoff & Liu, 1998; H. M. Yang & McConkie, 1999; G. Yan et al., 2006). Furthermore, I observed robust effects of the combinatorial properties of words. The frequency of a locative phrase influenced the probability of a fixation on the GROUND noun and the topological nominal, as well as the position and duration

of the initial fixation on both words. Fixation probabilities were lower, fixation durations were shorter, and fixation positions were more rightward for high frequency phrases as compared to low frequency phrases. In addition, I found effects of both the entropy and the relative entropy of the GROUND noun. In the context of locative phrases, entropy is a measure of the predictability of the topological nominal given the preposition and the GROUND noun, whereas relative entropy is a measure of the prototypicality of the frequency distribution of the nominal paradigm for a noun. Additional resources were allocated to both the GROUND noun and the nominal for locative phrases with high values for entropy and relative entropy. To my knowledge, the effects of phrase frequency, entropy, and relative entropy are the first phrase frequency, entropy, and relative entropy effects reported for simplified Chinese. The current findings fit well with previous results reported by Rayner et al. (2005) and H. C. Wang et al. (2010), who found effects of subjective predictability and transitional probability on eye movement patterns in sentence reading tasks. Not only the combinatorial properties of characters, but also the combinatorial properties of words thus influence lexical processing in simplified Chinese.

Both Chapter 3 and Chapter 4 reported the results of an analysis of the eye movement patterns during lexical processing. In addition to the lexical predictor effects discussed above, the eye tracking analyses in Chapter 3 and Chapter 4 revealed two interesting, more general aspects about lexical processing in simplified Chinese. First, characters and words are not processed independently. Instead, parafoveal preview allows for joint processing of the character or word that is fixated on and neighbouring characters or words, both on the right and on the left of the character or word that is fixated on. The importance of parafoveal preview for efficient processing is in line with previous findings (see, e.g., Inhoff & Liu, 1997; Inhoff, 1999; W. Liu et al., 2002; J. Yang et al., 2009; Tsai et al., 2004; M. Yan et al., 2009, 2012; Tsai et al., 2012).

Second, resources are dynamically allocated to information-rich areas. For the word naming task in Chapter 3, fixation durations of fixations on a character in two-character words were shorter when the character that was fixated on provided less information and when the other character provided more information. Conversely, fixation durations of fixations on a character were longer when the character that was fixated on provided more information and when the other character provided less information. Chapter 4 showed a similar pattern of results for lexical processing at the phrase level. Words that provided more information (i.e., low frequency

words and visually complex words) were fixated on more often than words that provided little information (i.e., high frequency words and visually simple words). Furthermore, fixation durations were longer and fixations were less far into the word for words that provided more information. The information provided by upcoming words played a role as well. When upcoming words contained more information, the fixation probability for the current word decreased and fixations on the current word were shorter and further into the word. The search for information thus proved a driving force behind eye fixation planning for words presented in isolation, as well as for words presented in phrasal and sentential contexts.

The lexical predictors in the CLD allowed for an in-depth analysis of the experimental data in Chapter 3 and Chapter 4, which provided important insights into lexical processing for simplified Chinese. Nonetheless, it should be noted that by no means I see the CLD as a completed project. There are at least two further types of lexical information I would like to add to the CLD in future research. First, the current version of the CLD does not contain semantic information. I plan to add both categorical and numerical semantic predictors to the CLD in the future. Categorical predictors could include information about the syntactic category of one-character and two-character words (e.g., noun, verb, et cetera), the syntactic composition of two-character words (e.g., noun-noun, verb-noun, et cetera), and the semantic structure of two-character words (e.g., modifier-head, head-modifier, et cetera). Recent advances in vector semantics (c.f., Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014) could help provide numerical estimates of the semantic similarity between a character and its semantic radical, between both characters in a two-character word, and between the characters in a two-character word and the two-character word as a whole. Furthermore, semantic neighbourhood density measures could be developed. The addition of semantic measures to the CLD would help to more efficiently address a wide range of research questions, including the ongoing single versus dual-route debate (i.e., "Is semantic access necessary to retrieve phonological forms from orthographic forms?"; see, e.g., Perfetti & Tan, 1998; Perfetti & Liu, 2006).

A second type of information that I would like to add to the CLD in the future is phonetic information. Currently, acoustic information in the CLD is described at a phonological level only. The auditory frequency, complexity and neighbourhood density measures are based on the segmental information provided by phonemes and diphones and the suprasegmental information provided by tones. It would be

interesting to provide information about the acoustic realization of characters in isolation and in the context of two-character words to the CLD. For this purpose, I plan to pre-process and publicly release the pronunciations of all 30,645 words from the single-participant study described in Chapter 3, along with descriptive statistics for these pronunciations (e.g., pitch contours) in the not too distant future.

Despite these shortcomings, the CLD contains a wealth of categorical and numerical information. I hope that the CLD will prove a valuable resource that helps improve the reliability, comparability and replicability of psycholinguistic research on simplified Chinese. In this dissertation, the application of the variables in the CLD in the analyses of the word naming data in Chapter 3 and the phrase reading data in Chapter 4 provided detailed insights into the quantitative and qualitative influence of the lexical information in the CLD on behavioural measures and eye movement patterns during lexical processing in Chinese. As noted by Baayen et al. (2016, p. 32), "exploratory data analysis is [particularly important] for those domains of inquiry where explicit and mathematically precise theories are lacking". The exploration of the influence of both categorical and lexical-distributional variables on lexical processing thus is an important first step towards a more comprehensive understanding of the language processing system for simplified Chinese. Where exploration ends, however, interpretation starts.

In my opinion, a thorough and objective examination of large-scale experimental data should precede any effort to understand the language processing system that underlies the effects of different lexical predictors. Hence, I mostly refrained from an interpretation of the results in terms of consequences for models of language processing in this dissertation. However, the results presented here provide important information about the nature of the language processing system for Chinese. One particularly striking finding in this dissertation was the abundance of effects related to the combinatorial properties of linguistic elements, both at the character level and at the word level. The leading model of lexical processing for Chinese – the multi-level interactive-activation model proposed by Taft and Zhu (1997) (see also Taft, 2006; Taft et al., 1999) – cannot straightforwardly account for the effects of entropy and relative entropy observed here. In a more general sense, verbal models may be too limited in their architectures and are not testable for the type of highly complex data reported here. Computationally implemented information-theoretic approaches to language processing, such as the naive discrimination learning framework proposed by Baayen et al. (2011) for language processing in English may be more promising.

The naive discrimination learning approach to language processing is based on a learning mechanism (Rescorla & Wagner, 1972) that is sensitive to the distributional properties of the language. For English, naive discrimination learning models have been shown to account for entropy effects on the reaction time in lexical decision (Baayen et al., 2011), as well as on the ERP signal in a picture naming task (Hendrix et al., 2016). The learning algorithm in naive discrimination learning models is a general learning mechanism that is aspecific to language and modality. In future research I hope to use the insights gained in this dissertation to explore the potential of the naive discrimination learning framework for language processing in simplified Chinese and the type of linguistic representations that should drive learning in this framework.

Appendices

A

Model summaries word naming

A.1 Naming latencies

A.1.1 One-character words

Table A.1: Model summary. Naming latencies for one-character words.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	-1.911	0.018	-104.569	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Session)	2.947	2.998	64.779	< 0.001
s(Initial Phoneme,bs="re")	23.738	29.000	14.208	< 0.001
s(PC1: C1 Frequency)	2.714	2.938	206.977	< 0.001
s(PC2: C1 Complexity)	2.028	2.452	37.262	< 0.001
s(PC11: C1 Traditional Frequency)	2.873	2.987	51.694	< 0.001

A.1.2 Two-character words

Table A.2: Model summary. Naming latencies for two-character words.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	-2.011	0.014	-143.122	< 0.001
C2 Type: Other	-0.013	0.005	-2.521	0.012
C2 Type: Pictologic	-0.031	0.011	-2.986	0.003
C2 Type: Pictosynthetic	-0.007	0.004	-1.960	0.050
C2 Type: Pictographic	-0.028	0.005	-5.414	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.963	2.999	156.845	< 0.001
s(Trial)	2.681	2.924	27.023	< 0.001
ti(Session, Trial)	4.249	5.294	6.292	< 0.001
s(Initial Phoneme,bs="re")	25.747	27.000	115.306	< 0.001
s(PC1: C1 Frequency)	2.853	2.983	162.057	< 0.001
s(PC2: C2 Frequency)	2.854	2.984	237.189	< 0.001
s(PC3: C2 Complexity)	1.000	1.000	58.266	< 0.001
s(PC4: C1 Complexity)	2.570	2.876	222.793	< 0.001
s(PC5: Frequency)	2.474	2.804	200.210	< 0.001
s(PC12: C1 Homographs)	2.962	2.998	18.880	< 0.001
s(PC16: C1 HLC Frequency)	2.325	2.703	13.757	< 0.001
s(PC22: C1 SR Frequency)	1.483	1.804	22.637	< 0.001
s(PC25: Frequency (SUBTL))	1.000	1.000	406.091	< 0.001
s(PC27: Phonological N)	1.000	1.000	24.677	< 0.001
s(PC28: Traditional Frequency)	2.176	2.576	10.867	< 0.001
s(PC34: C1 SR Complexity)	1.000	1.000	44.665	< 0.001
s(PC35: C1 RE)	2.887	2.989	10.610	< 0.001
s(PC36: Entropy Character Frequencies)	2.472	2.801	18.298	< 0.001
s(PC39: C2 Entropy)	1.562	1.915	41.597	< 0.001
s(PC40: C1 Traditional Frequency)	1.964	2.397	18.857	< 0.001
s(PC42: C1 Entropy)	1.000	1.000	134.006	< 0.001
s(PC50: C1 Trigram Entropy)	1.524	1.879	16.397	< 0.001
ti(PC1: C1 Freq., PC36: Entr. C. Freqs.)	3.073	3.271	13.377	< 0.001
ti(PC5: Frequency, PC35: C1 RE)	5.047	5.987	6.457	< 0.001

A.2 Pronunciation durations

A.2.1 One-character words

Table A.3: Model summary. Pronunciation durations for one-character words.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	308.979	6.869	44.980	< 0.001
C1 Tone: 2	31.660	1.859	17.034	< 0.001
C1 Tone: 3	77.753	2.019	38.510	< 0.001
C1 Tone: 4	-19.673	1.698	-11.588	< 0.001
C1 Tone: 5	26.167	8.098	3.231	0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	1.000	1.000	82.231	< 0.001
s(Trial)	2.755	2.954	16.359	< 0.001
s(Initial Phoneme,bs="re")	25.821	29.000	99.993	< 0.001
s(Final Phoneme,bs="re")	9.857	11.000	87.324	< 0.001
s(PC1: C1 Frequency)	2.381	2.736	18.090	< 0.001
s(PC6: C1 Diphone Frequency)	2.956	2.998	21.867	< 0.001
s(PC8: C1 Phoneme Frequency)	2.985	2.999	22.282	< 0.001
s(PC19: C1 Phonemes)	1.000	1.000	19.763	< 0.001

A.2.2 Two-character words

Table A.4: Model summary. Pronunciation durations for two-character words.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	475.971	8.480	56.125	< 0.001
C1 Tone: 2	1.947	0.700	2.780	0.005
C1 Tone: 3	4.422	0.709	6.235	< 0.001
C1 Tone: 4	3.457	0.630	5.491	< 0.001
C1 Tone: 5	5.512	12.674	0.435	0.664
C2 Tone: 2	18.423	0.739	24.916	< 0.001
C2 Tone: 3	-19.950	0.808	-24.697	< 0.001
C2 Tone: 4	-17.720	0.678	-26.118	< 0.001
C2 Tone: 5	-11.920	1.758	-6.783	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.986	3.000	2869.076	< 0.001
s(Trial)	2.944	2.998	227.724	< 0.001
ti(Session, Trial)	7.942	8.739	7.501	< 0.001
s(Initial Phoneme,bs="re")	26.454	27.000	324.999	< 0.001
s(Final Phoneme,bs="re")	11.754	12.000	196.077	< 0.001
s(PC5: Frequency)	2.281	2.654	65.601	< 0.001
s(PC6: C2 Phoneme Frequency)	2.978	2.999	135.753	< 0.001
s(PC7: C1 Diphone Frequency)	2.948	2.997	184.296	< 0.001
s(PC8: C2 Diphone Frequency)	2.929	2.996	26.589	< 0.001
s(PC9: C2 Homophones)	2.877	2.983	23.847	< 0.001
s(PC12: C1 Homographs)	2.890	2.991	12.934	< 0.001
s(PC14: C1 Min Phoneme Frequency)	2.792	2.970	9.661	< 0.001
s(PC17: C2 Min Phoneme Frequency)	1.000	1.000	443.405	< 0.001
s(PC18: C1 Phonological N)	2.777	2.963	56.981	< 0.001
s(PC19: C2 Phonological N)	2.618	2.893	28.008	< 0.001
s(PC20: Min Diphone Frequency)	2.933	2.997	46.367	< 0.001
s(PC25: Frequency (SUBTL))	2.075	2.487	53.320	< 0.001
s(PC38: C2 Phonemes)	2.882	2.988	37.857	< 0.001
s(PC43: C1 Phonemes)	2.897	2.991	69.885	< 0.001
s(PC45: C2 Picture Size)	2.848	2.983	11.918	< 0.001
ti(PC6: C2 Phon. Freq., PC9: C2 Homoph.)	12.024	13.730	11.551	< 0.001
ti(PC6: C2 Phon. Freq., PC20: Min Diph. Freq.)	7.274	8.787	9.017	< 0.001

A.3 Eye fixation durations

A.3.1 One-character words

Table A.5: Model summary. Eye fixation durations for one-character words: fixations that start from 400 to 200 ms before stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.238	0.016	390.864	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(X Position)	2.946	2.998	38.506	< 0.001

Table A.6: Model summary. Eye fixation durations for one-character words: fixations that start from 200 to 0 ms before stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.885	0.009	658.377	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Session)	2.803	2.971	12.644	< 0.001
s(X Position)	2.986	3.000	177.693	< 0.001

Table A.7: Model summary. Eye fixation durations for one-character words: fixations that start from 0 to 200 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.707	0.017	330.023	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(X Position)	2.966	2.999	196.122	< 0.001

Table A.8: Model summary. Eye fixation durations for one-character words: fixations that start from 200 to 400 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.306	0.011	579.218	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Trial)	1.000	1.000	18.397	< 0.001
s(X Position)	2.784	2.967	17.651	< 0.001
s(PC1: C1 Frequency)	2.226	2.608	32.994	< 0.001
s(PC2: C1 Complexity)	1.025	1.049	31.805	< 0.001

Table A.9: Model summary. Eye fixation durations for one-character words: fixations that start from 400 to 600 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.177	0.017	363.864	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(X Position)	2.861	2.985	31.736	< 0.001

Table A.10: Model summary. Eye fixation durations for one-character words: fixations that start from 600 to 800 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.072	0.015	418.415	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(X Position)	2.902	2.993	37.157	< 0.001
s(PC1: C1 Frequency)	1.000	1.000	29.556	< 0.001

Table A.11: Model summary. Eye fixation durations for one-character words: fixations that start from 800 to 1000 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.069	0.015	396.849	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(X Position)	2.925	2.996	39.482	< 0.001
s(PC1: C1 Frequency)	2.226	2.605	14.027	< 0.001

Table A.12: Model summary. Eye fixation durations for one-character words: fixations that start from 1000 to 1200 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.118	0.016	387.303	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(X Position)	2.927	2.996	51.205	< 0.001
s(PC1: C1 Frequency)	2.098	2.490	11.908	< 0.001

Table A.13: Model summary. Eye fixation durations for one-character words: fixations that start from 1200 to 1400 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.248	0.017	378.002	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(X Position)	2.892	2.992	31.190	< 0.001
s(PC1: C1 Frequency)	1.000	1.000	69.164	< 0.001

Table A.14: Model summary. Eye fixation durations for one-character words: fixations that start from 1400 to 1600 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.340	0.016	394.121	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(X Position)	2.886	2.990	31.274	< 0.001

A.3.2 Two-character words

Table A.15: Model summary. Eye fixation durations for two-character words: fixations that start from 400 to 200 ms before stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.187	0.006	1074.940	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(X Position)	2.976	3.000	282.906	< 0.001
s(PC4: C1 Complexity)	1.000	1.000	0.007	0.934
ti(PC4: C1 Complexity, X Position)	4.412	5.270	7.513	< 0.001

Table A.16: Model summary. Eye fixation durations for two-character words: fixations that start from 200 to 0 ms before stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.865	0.003	2113.659	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Session)	2.280	2.660	44.052	< 0.001
s(Trial)	2.709	2.936	4.439	0.010
s(X Position)	2.997	3.000	2789.001	< 0.001
s(Y Position)	2.768	2.961	14.547	< 0.001
ti(Session, X Position)	7.769	8.620	15.261	< 0.001
ti(Session, Trial)	2.846	2.982	19.315	< 0.001
ti(X Position, Y Position)	4.864	5.876	8.833	< 0.001
s(PC1: C1 Frequency)	2.364	2.737	6.099	0.004
s(PC2: C2 Frequency)	1.000	1.000	49.358	< 0.001
s(PC3: C2 Complexity)	1.962	2.383	31.736	< 0.001
s(PC4: C1 Complexity)	2.652	2.914	8.574	< 0.001
s(PC22: C1 SR Frequency)	1.000	1.000	22.026	< 0.001
s(PC36: Entropy Character Frequencies)	2.818	2.975	8.102	< 0.001
ti(PC1: C1 Frequency, X Position)	3.943	4.519	43.724	< 0.001
ti(PC2: C2 Frequency, X Position)	3.452	4.051	8.764	< 0.001
ti(PC3: C2 Complexity, X Position)	4.436	5.249	21.401	< 0.001
ti(PC4: C1 Complexity, X Position)	7.532	8.354	43.658	< 0.001
ti(PC1: C1 Frequency, PC4: C1 Complexity)	2.684	3.593	10.377	< 0.001

Table A.17: Model summary. Eye fixation durations for two-character words: fixations that start from 0 to 200 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.673	0.017	329.242	< 0.001
C1 Type: Pictologic	-0.124	0.027	-4.531	< 0.001
C1 Type: Pictophonetic	0.002	0.014	0.119	0.905
C1 Type: Pictosynthetic	-0.007	0.015	-0.501	0.617
C1 Type: Pictographic	-0.067	0.018	-3.645	< 0.001
C2 Type: Pictologic	0.089	0.029	3.063	0.002
C2 Type: Pictophonetic	0.013	0.014	0.979	0.328
C2 Type: Pictosynthetic	0.045	0.014	3.128	0.002
C2 Type: Pictographic	0.090	0.017	5.360	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.722	2.942	21.640	< 0.001
s(Trial)	2.702	2.933	10.569	< 0.001
s(X Position)	2.967	2.998	193.192	< 0.001
s(PC1: C1 Frequency)	2.381	2.744	128.258	< 0.001
s(PC2: C2 Frequency)	2.452	2.795	45.839	< 0.001
s(PC3: C2 Complexity)	2.717	2.945	51.568	< 0.001
s(PC4: C1 Complexity)	2.830	2.978	62.383	< 0.001
s(PC10: C1 Homophones)	1.459	1.771	21.120	< 0.001
s(PC23: C2 SR Frequency)	2.258	2.644	10.044	< 0.001
s(PC39: C2 Entropy)	2.849	2.983	10.325	< 0.001
s(PC42: C1 Entropy)	1.000	1.000	39.830	< 0.001
ti(PC1: C1 Frequency, X Position)	2.922	2.996	70.242	< 0.001
ti(PC3: C2 Complexity, X Position)	4.181	5.040	8.931	< 0.001
ti(PC4: C1 Complexity, X Position)	4.628	5.347	36.178	< 0.001
ti(PC1: C1 Frequency, PC4: C1 Complexity)	1.872	2.530	36.414	< 0.001

Table A.18: Model summary. Eye fixation durations for two-character words: fixations that start from 200 to 400 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.751	0.007	828.825	< 0.001
C2 Type: Other	0.038	0.015	2.551	0.011
C2 Type: Pictologic	0.126	0.033	3.782	< 0.001
C2 Type: Pictosynthetic	0.024	0.011	2.184	0.029
C2 Type: Pictographic	0.088	0.017	5.357	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.779	2.962	10.789	< 0.001
s(Trial)	2.318	2.684	10.792	< 0.001
s(X Position)	2.993	3.000	822.123	< 0.001
ti(Session, X Position)	5.900	7.058	7.111	< 0.001
s(PC1: C1 Frequency)	1.000	1.001	0.012	0.911
s(PC2: C2 Frequency)	2.514	2.843	26.091	< 0.001
s(PC3: C2 Complexity)	1.000	1.000	12.538	< 0.001
s(PC4: C1 Complexity)	1.000	1.000	7.816	0.005
s(PC5: Frequency)	1.000	1.000	11.896	< 0.001
s(PC23: C2 SR Frequency)	2.189	2.583	2.769	0.094
s(PC32: C2 RE)	1.000	1.000	21.116	< 0.001
s(PC42: C1 Entropy)	1.000	1.000	5.279	0.022
ti(PC1: C1 Frequency, X Position)	6.230	7.149	42.349	< 0.001
ti(PC2: C2 Frequency, X Position)	3.045	3.568	31.669	< 0.001
ti(PC3: C2 Complexity, X Position)	5.874	6.901	28.593	< 0.001
ti(PC4: C1 Complexity, X Position)	5.453	6.364	25.165	< 0.001
ti(PC5: Frequency, X Position)	2.939	3.132	12.614	< 0.001
ti(PC23: C2 SR Frequency, X Position)	4.654	5.796	7.119	< 0.001
ti(PC42: C1 Entropy, X Position)	3.386	4.428	11.681	< 0.001

Table A.19: Model summary. Eye fixation durations for two-character words: fixations that start from 400 to 600 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.012	0.018	332.218	< 0.001
C1 Tone: 2	-0.017	0.013	-1.313	0.189
C1 Tone: 3	-0.077	0.014	-5.665	< 0.001
C1 Tone: 4	-0.029	0.012	-2.466	0.014
C1 Tone: 5	0.043	0.031	1.390	0.165
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.915	2.994	52.451	< 0.001
s(Final Phoneme,bs="re")	8.687	12.000	12.460	< 0.001
s(X Position)	2.993	3.000	272.754	< 0.001
s(Y Position)	1.633	2.006	16.551	< 0.001
ti(X Position, Y Position)	4.187	5.295	6.258	< 0.001
s(PC1: C1 Frequency)	1.000	1.000	16.757	< 0.001
s(PC2: C2 Frequency)	2.561	2.866	5.726	0.004
s(PC3: C2 Complexity)	2.481	2.825	18.018	< 0.001
s(PC4: C1 Complexity)	2.602	2.894	16.552	< 0.001
s(PC8: C2 Diphone Frequency)	2.585	2.870	11.002	< 0.001
s(PC13: C2 Homographs)	1.652	1.978	0.239	0.770
s(PC22: C1 SR Frequency)	2.306	2.683	1.235	0.233
s(PC36: Entropy Character Frequencies)	1.000	1.000	6.451	0.011
ti(PC1: C1 Frequency, X Position)	1.159	1.299	43.084	< 0.001
ti(PC2: C2 Frequency, X Position)	6.218	7.258	10.993	< 0.001
ti(PC3: C2 Complexity, X Position)	2.867	2.986	20.911	< 0.001
ti(PC13: C2 Homographs, X Position)	5.284	6.135	5.696	< 0.001
ti(PC22: C1 SR Frequency, X Position)	2.603	3.417	7.661	< 0.001
ti(PC36: Entropy Character Frequencies, X Position)	2.361	3.001	8.797	< 0.001

Table A.20: Model summary. Eye fixation durations for two-character words: fixations that start from 600 to 800 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.992	0.005	1088.370	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.920	2.995	41.653	< 0.001
s(X Position)	2.586	2.883	39.976	< 0.001
s(PC1: C1 Frequency)	1.000	1.000	25.269	< 0.001
s(PC2: C2 Frequency)	1.000	1.000	27.809	< 0.001
s(PC5: Frequency)	2.007	2.400	12.006	< 0.001

Table A.21: Model summary. Eye fixation durations for two-character words: fixations that start from 800 to 1000 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.017	0.005	1149.838	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.909	2.993	40.239	< 0.001
s(X Position)	2.808	2.973	170.881	< 0.001
s(PC2: C2 Frequency)	1.741	2.129	10.991	< 0.001
s(PC3: C2 Complexity)	1.000	1.000	20.526	< 0.001
s(PC4: C1 Complexity)	1.000	1.000	5.425	0.020
s(PC5: Frequency)	2.170	2.558	13.429	< 0.001
ti(PC4: C1 Complexity, X Position)	5.232	6.222	10.373	< 0.001

Table A.22: Model summary. Eye fixation durations for two-character words: fixations that start from 1000 to 1200 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.033	0.006	1047.318	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Session)	2.798	2.969	12.871	< 0.001
s(X Position)	2.973	2.999	172.551	< 0.001
s(PC5: Frequency)	2.068	2.460	13.108	< 0.001

Table A.23: Model summary. Eye fixation durations for two-character words: fixations that start from 1200 to 1400 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.086	0.006	1010.116	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.948	2.998	41.467	< 0.001
s(X Position)	2.966	2.999	150.722	< 0.001
ti(Session, X Position)	6.054	7.318	6.530	< 0.001
s(PC1: C1 Frequency)	1.526	1.862	19.742	< 0.001
s(PC4: C1 Complexity)	1.000	1.000	26.210	< 0.001
s(PC5: Frequency)	1.704	2.075	15.650	< 0.001

Table A.24: Model summary. Eye fixation durations for two-character words: fixations that start from 1400 to 1600 ms after stimulus onset.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	6.189	0.006	1036.706	< 0.001
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Session)	2.931	2.996	30.151	< 0.001
s(X Position)	2.948	2.998	141.358	< 0.001
ti(Session, X Position)	2.695	2.930	10.578	< 0.001
s(PC1: C1 Frequency)	1.562	1.909	20.638	< 0.001
s(PC4: C1 Complexity)	1.731	2.130	11.997	< 0.001

B

Model summaries phrase reading

B.1 Experiment 1

B.1.1 Preposition

B.1.1.1 Probability

Table B.1: Model summary. Fixation Probability for the preposition in Experiment 1.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	0.584	0.134	4.357	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	23.189	24.000	631.743	< 0.001
s(Trial)	2.280	2.652	14.738	0.003
s(PC1: Topological Nominal Frequency)	1.000	1.000	4.673	0.031

Table B.2: GBM variable importance. Fixation Probability for the preposition in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (8.26%).

rank	predictor	rel. infl	abs. infl.
1	Participant	53.34	6.36
2	Item	23.89	2.85
3	Trial	12.52	1.49
4	PC2: GROUND Noun Frequency	1.59	0.19
5	PC1: Topological Nominal Frequency	1.24	0.15
6	PC4: Phrase Frequency	1.07	0.13
7	PC11: Topological Nominal Length	1.05	0.12
8	PC5: RE	0.92	0.11
9	PC7: Topological Nominal Visual Complexity	0.81	0.10
10	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.81	0.10
11	PC8: GROUND Noun Length	0.72	0.09
12	PC6: Entropy	0.72	0.09
13	PC10: GROUND Noun Picture Size	0.72	0.09
14	PC3: GROUND Noun Visual Complexity	0.61	0.07

B.1.1.2 Position

Table B.3: Model summary. Fixation Position for initial fixations on the preposition in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	16.186	0.821	19.727	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	22.911	24.000	22.732	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	4.583	0.032

Table B.4: GBM variable importance. Fixation Position for initial fixations on the preposition in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (8.85%).

rank	predictor	rel. infl	abs. infl.
1	Participant	46.88	6.76
2	Item	23.11	3.33
3	Trial	11.62	1.68
4	Y Position	6.09	0.88
5	PC2: GROUND Noun Frequency	1.87	0.27
6	PC4: Phrase Frequency	1.44	0.21
7	PC6: Entropy	1.35	0.19
8	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.32	0.19
9	PC1: Topological Nominal Frequency	1.15	0.17
10	PC5: RE	1.08	0.16
11	PC8: GROUND Noun Length	1.05	0.15
12	PC7: Topological Nominal Visual Complexity	1.02	0.15
13	PC11: Topological Nominal Length	0.93	0.13
14	PC3: GROUND Noun Visual Complexity	0.71	0.10
15	PC10: GROUND Noun Picture Size	0.38	0.06

B.1.1.3 Duration

Table B.5: Model summary. Fixation Duration for initial fixations on the preposition in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.223	0.020	263.563	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	23.544	24.000	47.762	< 0.001
s(X Position)	2.513	2.829	59.797	< 0.001
s(PC1: Topological Nominal Frequency)	1.000	1.000	14.261	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.615	1.964	17.805	< 0.001
s(PC10: GROUND Noun Picture Size)	2.279	2.673	2.516	0.043

Table B.6: GBM variable importance. Fixation Duration for initial fixations on the preposition in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (19.49%).

rank	predictor	rel. infl	abs. infl.
1	Participant	35.77	8.91
2	Item	21.35	5.32
3	X Position	13.56	3.38
4	Trial	7.88	1.96
5	Y Position	7.72	1.92
6	PC8: GROUND Noun Length	2.53	0.63
7	PC1: Topological Nominal Frequency	1.86	0.46
8	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.76	0.44
9	PC6: Entropy	1.45	0.36
10	PC5: RE	1.32	0.33
11	PC11: Topological Nominal Length	1.15	0.29
12	PC2: GROUND Noun Frequency	0.92	0.23
13	PC10: GROUND Noun Picture Size	0.85	0.21
14	PC3: GROUND Noun Visual Complexity	0.77	0.19
15	PC7: Topological Nominal Visual Complexity	0.67	0.17
16	PC4: Phrase Frequency	0.47	0.12

B.1.2 GROUND noun

B.1.2.1 Probability

Table B.7: Model summary. Fixation Probability for the GROUND noun in Experiment 1.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	3.349	0.265	12.661	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	22.189	24.000	376.471	< 0.001
s(Item,bs="re")	33.551	296.000	40.539	0.015
s(Trial)	2.846	2.980	16.829	< 0.001
s(PC1: Topological Nominal Frequency)	1.000	1.000	43.053	< 0.001
s(PC2: GROUND Noun Frequency)	1.000	1.000	22.814	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.333	1.547	46.384	< 0.001
s(PC7: Top. Nominal Visual Complexity)	1.000	1.000	7.519	0.006
s(PC8: GROUND Noun Length)	1.000	1.000	66.924	< 0.001
s(PC11: Topological Nominal Length)	1.000	1.000	13.198	< 0.001

Table B.8: GBM variable importance. Fixation Probability for the GROUND noun in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (12.56%).

rank	predictor	rel. infl	abs. infl.
1	Item	48.50	10.09
2	Participant	35.57	7.40
3	Trial	7.18	1.49
4	PC8: GROUND Noun Length	4.12	0.86
5	PC1: Topological Nominal Frequency	1.17	0.24
6	PC11: Topological Nominal Length	0.73	0.15
7	PC2: GROUND Noun Frequency	0.69	0.14
8	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.40	0.08
9	PC5: RE	0.36	0.08
10	PC7: Topological Nominal Visual Complexity	0.33	0.07
11	PC10: GROUND Noun Picture Size	0.32	0.07
12	PC6: Entropy	0.26	0.05
13	PC3: GROUND Noun Visual Complexity	0.26	0.05
14	PC4: Phrase Frequency	0.10	0.02

B.1.2.2 Position

Table B.9: Model summary. Fixation Position for initial fixations on the GROUND noun in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	43.827	2.520	17.389	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	23.882	24.000	177.010	< 0.001
s(Item,bs="re")	97.228	288.000	0.534	< 0.001
s(Trial)	2.656	2.910	4.115	0.004
s(Previous Fixation Duration)	2.980	2.999	121.614	< 0.001
s(Partial Saccade Length)	2.829	2.977	938.628	< 0.001
s(PC1: Topological Nominal Frequency)	1.485	1.692	426.392	< 0.001
s(PC2: GROUND Noun Frequency)	1.000	1.000	10.518	0.001
s(PC3: GROUND Noun Visual Complexity)	2.823	2.913	27.071	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	107.037	< 0.001
s(PC6: Entropy)	2.858	2.940	6.317	< 0.001
s(PC7: Top. Nominal Visual Complexity)	1.000	1.000	77.203	< 0.001
s(PC8: GROUND Noun Length)	2.353	2.594	174.608	< 0.001
s(PC10: GROUND Noun Picture Size)	1.000	1.000	19.128	< 0.001
s(PC11: Topological Nominal Length)	1.000	1.000	114.506	< 0.001

Table B.10: GBM variable importance. Fixation Position for initial fixations on the GROUND noun in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (55.79%).

rank	predictor	rel. infl	abs. infl.
1	Participant	34.36	19.95
2	Item	33.50	19.45
3	Partial Saccade Length	23.71	13.77
4	Previous Fixation Duration	3.75	2.18
5	PC8: GROUND Noun Length	1.33	0.77
6	Trial	1.08	0.63
7	PC1: Topological Nominal Frequency	0.81	0.47
8	Y Position	0.39	0.23
9	PC11: Topological Nominal Length	0.27	0.16
10	PC6: Entropy	0.20	0.11
11	PC7: Topological Nominal Visual Complexity	0.16	0.09
12	PC4: Phrase Frequency	0.14	0.08
13	PC5: RE	0.09	0.05
14	PC3: GROUND Noun Visual Complexity	0.08	0.05
15	PC10: GROUND Noun Picture Size	0.07	0.04
16	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.03	0.02
17	PC2: GROUND Noun Frequency	0.02	0.01

B.1.2.3 Duration

Table B.11: Model summary. Fixation Duration for initial fixations on the GROUND noun in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.354	0.046	115.082	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	23.774	24.000	104.955	< 0.001
s(Item,bs="re")	64.578	310.000	0.267	0.002
s(Trial)	1.520	1.845	135.663	< 0.001
s(X Position)	1.294	1.522	10.211	< 0.001
s(Y Position)	1.000	1.000	3.862	0.049
s(Previous Fixation Duration)	1.000	1.000	7.061	0.008
s(Saccade Length)	1.000	1.000	4.612	0.032
s(PC8: GROUND Noun Length)	1.221	1.368	11.761	< 0.001

Table B.12: GBM variable importance. Fixation Duration for initial fixations on the GROUND noun in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (30.25%).

rank	predictor	rel. infl	abs. infl.
1	Participant	66.87	23.48
2	Trial	8.92	3.13
3	Item	8.65	3.04
4	Previous Fixation Duration	4.00	1.41
5	Saccade Length	2.50	0.88
6	X Position	2.06	0.72
7	PC1: Topological Nominal Frequency	1.32	0.46
8	Y Position	1.19	0.42
9	PC8: GROUND Noun Length	0.93	0.33
10	PC5: RE	0.78	0.27
11	PC3: GROUND Noun Visual Complexity	0.67	0.23
12	PC7: Topological Nominal Visual Complexity	0.59	0.21
13	PC10: GROUND Noun Picture Size	0.49	0.17
14	PC2: GROUND Noun Frequency	0.28	0.10
15	PC11: Topological Nominal Length	0.24	0.08
16	PC6: Entropy	0.23	0.08
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.15	0.05
18	PC4: Phrase Frequency	0.12	0.04

B.1.3 Nominal

B.1.3.1 Probability

Table B.13: Model summary. Fixation Probability for the topological nominal in Experiment 1.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	-0.950	0.200	-4.746	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	23.458	24.000	844.655	< 0.001
s(Item,bs="re")	61.454	305.000	79.506	< 0.001
s(Trial)	1.000	1.000	73.191	< 0.001
s(PC1: Topological Nominal Frequency)	1.000	1.001	880.466	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.990	2.298	34.368	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	122.627	< 0.001
s(PC7: Top. Nominal Visual Complexity)	1.052	1.091	40.259	< 0.001
s(PC11: Topological Nominal Length)	2.709	2.889	102.773	< 0.001

Table B.14: GBM variable importance. Fixation Probability for the topological nominal in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (31.30%).

rank	predictor	rel. infl	abs. infl.
1	Item	59.15	20.87
2	Participant	31.41	11.08
3	Trial	6.20	2.19
4	PC3: GROUND Noun Visual Complexity	0.48	0.17
5	PC8: GROUND Noun Length	0.48	0.17
6	PC7: Topological Nominal Visual Complexity	0.47	0.17
7	PC10: GROUND Noun Picture Size	0.46	0.16
8	PC5: RE	0.36	0.13
9	PC1: Topological Nominal Frequency	0.32	0.11
10	PC2: GROUND Noun Frequency	0.20	0.07
11	PC6: Entropy	0.18	0.06
12	PC4: Phrase Frequency	0.18	0.06
13	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.08	0.03
14	PC11: Topological Nominal Length	0.04	0.01

B.1.3.2 Position

Table B.15: Model summary. Fixation Position for initial fixations on the topological nominal in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	22.044	1.163	18.953	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	22.246	24.000	19.313	< 0.001
s(Item,bs="re")	29.409	290.000	0.121	0.037
s(Trial)	1.000	1.000	13.139	< 0.001
s(Y Position)	2.363	2.737	8.505	< 0.001
s(Previous Fixation Duration)	2.582	2.880	7.984	< 0.001
s(Partial Saccade Length)	2.878	2.987	246.763	< 0.001
s(PC1: Topological Nominal Frequency)	2.047	2.332	33.734	< 0.001
s(PC3: GROUND Noun Visual Complexity)	2.708	2.890	3.521	0.033
s(PC4: Phrase Frequency)	1.000	1.000	5.064	0.025
s(PC7: Top. Nominal Visual Complexity)	2.740	2.915	5.064	0.004
s(PC8: GROUND Noun Length)	2.321	2.602	5.107	0.002
s(PC11: Topological Nominal Length)	2.175	2.495	54.087	< 0.001

Table B.16: GBM variable importance. Fixation Position for initial fixations on the topological nominal in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (16.07%).

rank	predictor	rel. infl	abs. infl.
1	Item	32.89	8.74
2	Participant	29.62	7.87
3	Trial	9.49	2.52
4	Partial Saccade Length	4.87	1.30
5	Y Position	4.85	1.29
6	Previous Fixation Duration	4.54	1.21
7	PC1: Topological Nominal Frequency	2.53	0.67
8	PC11: Topological Nominal Length	1.82	0.48
9	PC5: RE	1.62	0.43
10	PC3: GROUND Noun Visual Complexity	1.52	0.40
11	PC4: Phrase Frequency	1.35	0.36
12	PC8: GROUND Noun Length	1.14	0.30
13	PC7: Topological Nominal Visual Complexity	1.01	0.27
14	PC10: GROUND Noun Picture Size	1.00	0.27
15	PC2: GROUND Noun Frequency	0.96	0.26
16	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.59	0.16
17	PC6: Entropy	0.19	0.05

B.1.3.3 Duration

Table B.17: Model summary. Fixation Duration for initial fixations on the topological nominal in Experiment 1.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.114	0.043	118.392	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	22.813	24.000	17.946	< 0.001
s(Trial)	2.544	2.850	23.797	< 0.001
s(X Position)	1.000	1.000	143.773	< 0.001
s(Y Position)	1.000	1.000	10.091	0.002
s(Previous Fixation Duration)	2.399	2.770	2.994	0.021
s(PC1: Topological Nominal Frequency)	1.080	1.151	106.650	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	8.615	0.003
s(PC8: GROUND Noun Length)	2.255	2.631	4.386	0.005
s(PC11: Topological Nominal Length)	1.300	1.529	21.715	< 0.001

Table B.18: GBM variable importance. Fixation Duration for initial fixations on the topological nominal in Experiment 1. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (19.37%).

rank	predictor	rel. infl	abs. infl.
1	Item	29.32	8.69
2	Participant	27.47	8.14
3	X Position	11.19	3.32
4	Trial	7.80	2.31
5	Saccade Length	4.04	1.20
6	PC1: Topological Nominal Frequency	3.39	1.01
7	Previous Fixation Duration	3.23	0.96
8	Y Position	3.07	0.91
9	PC3: GROUND Noun Visual Complexity	1.62	0.48
10	PC5: RE	1.42	0.42
11	PC11: Topological Nominal Length	1.38	0.41
12	PC8: GROUND Noun Length	1.08	0.32
13	PC10: GROUND Noun Picture Size	1.05	0.31
14	PC4: Phrase Frequency	1.02	0.30
15	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.92	0.27
16	PC6: Entropy	0.89	0.26
17	PC2: GROUND Noun Frequency	0.80	0.24
18	PC7: Topological Nominal Visual Complexity	0.32	0.10

B.2 Experiment 2: Early Locative sentences

B.2.1 Preposition

B.2.1.1 Probability

Table B.19: Model summary. Fixation Probability for the preposition for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	0.375	0.137	2.748	0.006
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	27.422	29.000	424.409	< 0.001
s(Item,bs="re")	37.793	155.000	50.200	0.004
s(Trial)	1.000	1.000	19.121	< 0.001
s(PC8: GROUND Noun Length)	1.000	1.000	4.778	0.029

Table B.20: GBM variable importance. Fixation Probability for the preposition for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (7.62%).

rank	predictor	rel. infl	abs. infl.
1	Participant	53.31	6.63
2	Item	22.11	2.75
3	Trial	12.99	1.62
4	PC1: Topological Nominal Frequency	2.52	0.31
5	PC4: Phrase Frequency	1.52	0.19
6	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.36	0.17
7	PC3: GROUND Noun Visual Complexity	1.06	0.13
8	PC5: RE	0.86	0.11
9	PC6: Entropy	0.83	0.10
10	PC2: GROUND Noun Frequency	0.76	0.09
11	PC8: GROUND Noun Length	0.70	0.09
12	PC11: Topological Nominal Length	0.69	0.09
13	PC7: Topological Nominal Visual Complexity	0.68	0.08
14	PC10: GROUND Noun Picture Size	0.61	0.08

B.2.1.2 Position

Table B.21: Model summary. Fixation Position for initial fixations on the preposition for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	19.353	0.530	36.532	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	21.490	29.000	3.192	< 0.001
s(Trial)	1.900	2.288	3.940	0.016
s(Partial Saccade Length)	2.781	2.963	88.463	< 0.001

Table B.22: GBM variable importance. Fixation Position for initial fixations on the preposition for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (6.79%).

rank	predictor	rel. infl	abs. infl.
1	Partial Saccade Length	33.31	4.94
2	Item	21.16	3.14
3	Participant	13.32	1.98
4	Previous Fixation Duration	8.21	1.22
5	Trial	6.73	1.00
6	Y Position	4.81	0.71
7	PC1: Topological Nominal Frequency	1.89	0.28
8	PC2: GROUND Noun Frequency	1.52	0.23
9	PC5: RE	1.52	0.23
10	PC11: Topological Nominal Length	1.30	0.19
11	PC10: GROUND Noun Picture Size	1.07	0.16
12	PC8: GROUND Noun Length	0.98	0.15
13	PC7: Topological Nominal Visual Complexity	0.94	0.14
14	PC3: GROUND Noun Visual Complexity	0.92	0.14
15	PC6: Entropy	0.88	0.13
16	PC4: Phrase Frequency	0.73	0.11
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.68	0.10

B.2.1.3 Duration

Table B.23: Model summary. Fixation Duration for initial fixations on the preposition for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.258	0.034	154.560	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	26.277	29.000	17.872	< 0.001
s(Previous Fixation Duration)	1.000	1.000	4.317	0.038
s(Saccade Length)	2.712	2.940	17.691	< 0.001
s(PC8: GROUND Noun Length)	2.325	2.690	5.939	< 0.001
s(PC11: Topological Nominal Length)	2.229	2.634	3.326	0.021

Table B.24: GBM variable importance. Fixation Duration for initial fixations on the preposition for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (20.09%).

rank	predictor	rel. infl	abs. infl.
1	Participant	41.33	10.57
2	Item	13.97	3.58
3	Saccade Length	11.54	2.95
4	Trial	6.99	1.79
5	Previous Fixation Duration	5.69	1.46
6	X Position	4.97	1.27
7	Y Position	4.79	1.23
8	PC10: GROUND Noun Picture Size	1.59	0.41
9	PC2: GROUND Noun Frequency	1.34	0.34
10	PC3: GROUND Noun Visual Complexity	1.32	0.34
11	PC5: RE	1.05	0.27
12	PC4: Phrase Frequency	1.01	0.26
13	PC7: Topological Nominal Visual Complexity	0.96	0.25
14	PC6: Entropy	0.94	0.24
15	PC11: Topological Nominal Length	0.89	0.23
16	PC1: Topological Nominal Frequency	0.68	0.17
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.54	0.14
18	PC8: GROUND Noun Length	0.40	0.10

B.2.2 GROUND noun

B.2.2.1 Probability

Table B.25: Model summary. Fixation Probability for the GROUND noun for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	3.112	0.199	15.662	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	24.788	29.000	301.512	< 0.001
s(Item,bs="re")	25.598	148.000	32.831	0.019
s(PC2: GROUND Noun Frequency)	1.000	1.000	21.142	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	30.742	< 0.001
s(PC8: GROUND Noun Length)	1.000	1.000	26.124	< 0.001
s(PC9: GROUND Noun Freq. (SUBTLEX))	1.000	1.000	13.159	< 0.001
s(PC10: GROUND Noun Picture Size)	1.000	1.000	4.511	0.034

Table B.26: GBM variable importance. Fixation Probability for the GROUND noun for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (12.22%).

rank	predictor	rel. infl	abs. infl.
1	Item	45.42	9.00
2	Participant	37.85	7.50
3	Trial	11.29	2.24
4	PC8: GROUND Noun Length	1.35	0.27
5	PC3: GROUND Noun Visual Complexity	1.07	0.21
6	PC5: RE	0.99	0.20
7	PC7: Topological Nominal Visual Complexity	0.83	0.16
8	PC4: Phrase Frequency	0.41	0.08
9	PC1: Topological Nominal Frequency	0.24	0.05
10	PC2: GROUND Noun Frequency	0.21	0.04
11	PC6: Entropy	0.17	0.03
12	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.08	0.02
13	PC11: Topological Nominal Length	0.07	0.01
14	PC10: GROUND Noun Picture Size	0.02	0.00

B.2.2.2 Position

Table B.27: Model summary. Fixation Position for initial fixations on the GROUND noun for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	17.182	4.101	4.189	< 0.001
LengthNoun	6.041	2.025	2.983	0.003
smooth terms	**edf**	**ref. df**	**F-value**	**p-value**
s(Participant,bs="re")	27.872	29.000	18.843	< 0.001
s(Previous Fixation Duration)	1.000	1.000	9.248	0.002
s(Partial Saccade Length)	2.968	2.999	404.073	< 0.001
s(PC2: GROUND Noun Frequency)	1.000	1.000	6.554	0.011
s(PC5: RE)	1.000	1.000	3.903	0.048
s(PC7: Top. Nominal Visual Complexity)	1.158	1.298	5.035	0.026
s(PC8: GROUND Noun Length)	2.097	2.554	5.293	0.002

Table B.28: GBM variable importance. Fixation Position for initial fixations on the GROUND noun for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (27.60%).

rank	predictor	rel. infl	abs. infl.
1	Partial Saccade Length	45.51	15.09
2	Participant	26.64	8.83
3	Item	13.65	4.52
4	Previous Fixation Duration	3.98	1.32
5	Y Position	2.49	0.83
6	PC8: GROUND Noun Length	1.63	0.54
7	Trial	1.45	0.48
8	PC2: GROUND Noun Frequency	0.83	0.28
9	PC10: GROUND Noun Picture Size	0.59	0.20
10	PC11: Topological Nominal Length	0.50	0.17
11	PC7: Topological Nominal Visual Complexity	0.50	0.17
12	PC6: Entropy	0.47	0.15
13	PC5: RE	0.46	0.15
14	PC1: Topological Nominal Frequency	0.42	0.14
15	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.36	0.12
16	PC3: GROUND Noun Visual Complexity	0.28	0.09
17	PC4: Phrase Frequency	0.23	0.08

B.2.2.3 Duration

Table B.29: Model summary. Fixation Duration for initial fixations on the GROUND noun for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.340	0.034	156.017	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	28.051	29.000	29.299	< 0.001
s(Trial)	2.703	2.934	5.382	< 0.001
s(PC3: GROUND Noun Visual Complexity)	2.847	2.982	5.326	0.001
s(PC8: GROUND Noun Length)	1.000	1.000	13.472	< 0.001

Table B.30: GBM variable importance. Fixation Duration for initial fixations on the GROUND noun for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (13.60%).

rank	predictor	rel. infl	abs. infl.
1	Participant	51.88	10.30
2	Item	10.39	2.06
3	Saccade Length	8.80	1.75
4	Y Position	5.62	1.12
5	X Position	4.84	0.96
6	Previous Fixation Duration	4.48	0.89
7	Trial	4.08	0.81
8	PC5: RE	1.49	0.30
9	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.25	0.25
10	PC10: GROUND Noun Picture Size	1.14	0.23
11	PC3: GROUND Noun Visual Complexity	1.10	0.22
12	PC4: Phrase Frequency	1.05	0.21
13	PC1: Topological Nominal Frequency	0.93	0.18
14	PC8: GROUND Noun Length	0.77	0.15
15	PC11: Topological Nominal Length	0.71	0.14
16	PC6: Entropy	0.50	0.10
17	PC2: GROUND Noun Frequency	0.49	0.10
18	PC7: Topological Nominal Visual Complexity	0.47	0.09

B.2.3 Nominal

B.2.3.1 Probability

Table B.31: Model summary. Fixation Probability for the topological nominal for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	1.758	0.120	14.644	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	25.345	29.000	212.085	< 0.001
s(Trial)	1.000	1.000	12.228	< 0.001
s(PC1: Topological Nominal Frequency)	1.009	1.018	268.201	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	9.770	0.002
s(PC4: Phrase Frequency)	1.670	2.042	37.788	< 0.001
s(PC6: Entropy)	1.000	1.000	4.345	0.037
s(PC8: GROUND Noun Length)	1.000	1.000	4.170	0.041
s(PC11: Topological Nominal Length)	1.000	1.000	14.892	< 0.001

Table B.32: GBM variable importance. Fixation Probability for the topological nominal for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (9.91%).

rank	predictor	rel. infl	abs. infl.
1	Item	52.85	7.94
2	Participant	25.56	3.84
3	Trial	12.04	1.81
4	PC1: Topological Nominal Frequency	2.96	0.44
5	PC5: RE	1.07	0.16
6	PC7: Topological Nominal Visual Complexity	1.04	0.16
7	PC10: GROUND Noun Picture Size	0.94	0.14
8	PC11: Topological Nominal Length	0.82	0.12
9	PC6: Entropy	0.69	0.10
10	PC4: Phrase Frequency	0.67	0.10
11	PC2: GROUND Noun Frequency	0.48	0.07
12	PC8: GROUND Noun Length	0.39	0.06
13	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.26	0.04
14	PC3: GROUND Noun Visual Complexity	0.21	0.03

B.2.3.2 Position

Table B.33: Model summary. Fixation Position for initial fixations on the topological nominal for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	25.556	0.949	26.939	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	27.025	29.000	11.989	< 0.001
s(Item,bs="re")	49.956	142.000	0.592	< 0.001
s(Previous Fixation Duration)	2.490	2.827	5.655	< 0.001
s(Partial Saccade Length)	2.947	2.997	354.798	< 0.001
s(PC1: Topological Nominal Frequency)	2.658	2.815	18.809	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	14.821	< 0.001
s(PC7: Top. Nominal Visual Complexity)	2.752	2.879	2.471	0.039
s(PC8: GROUND Noun Length)	1.000	1.000	13.613	< 0.001
s(PC9: GROUND Noun Freq. (SUBTLEX))	1.000	1.000	12.549	< 0.001
s(PC10: GROUND Noun Picture Size)	2.125	2.386	3.826	0.025
s(PC11: Topological Nominal Length)	1.000	1.000	129.906	< 0.001

Table B.34: GBM variable importance. Fixation Position for initial fixations on the topological nominal for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (34.32%).

rank	predictor	rel. infl	abs. infl.
1	Partial Saccade Length	33.31	13.32
2	Item	21.16	8.46
3	Participant	13.32	5.33
4	Previous Fixation Duration	8.21	3.28
5	Trial	6.73	2.69
6	Y Position	4.81	1.92
7	PC1: Topological Nominal Frequency	1.89	0.76
8	PC2: GROUND Noun Frequency	1.52	0.61
9	PC5: RE	1.52	0.61
10	PC11: Topological Nominal Length	1.30	0.52
11	PC10: GROUND Noun Picture Size	1.07	0.43
12	PC8: GROUND Noun Length	0.98	0.39
13	PC7: Topological Nominal Visual Complexity	0.94	0.38
14	PC3: GROUND Noun Visual Complexity	0.92	0.37
15	PC6: Entropy	0.88	0.35
16	PC4: Phrase Frequency	0.73	0.29
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.68	0.27

B.2.3.3 Duration

Table B.35: Model summary. Fixation Duration for initial fixations on the topological nominal for Early Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.338	0.029	184.437	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	27.407	29.000	16.094	< 0.001
s(Item,bs="re")	25.263	152.000	0.202	0.050
s(Previous Fixation Duration)	2.239	2.644	3.849	0.011
s(Saccade Length)	2.647	2.907	9.006	< 0.001
s(PC1: Topological Nominal Frequency)	1.000	1.000	20.817	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	16.080	< 0.001
s(PC11: Topological Nominal Length)	1.000	1.000	16.067	< 0.001

Table B.36: GBM variable importance. Fixation Duration for initial fixations on the topological nominal for Early Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (10.91%).

rank	predictor	rel. infl	abs. infl.
1	Participant	36.91	6.68
2	Item	17.92	3.24
3	Saccade Length	11.24	2.03
4	Previous Fixation Duration	6.84	1.24
5	Y Position	6.35	1.15
6	Trial	4.29	0.78
7	X Position	2.80	0.51
8	PC3: GROUND Noun Visual Complexity	2.31	0.42
9	PC11: Topological Nominal Length	2.09	0.38
10	PC8: GROUND Noun Length	1.69	0.31
11	PC7: Topological Nominal Visual Complexity	1.59	0.29
12	PC2: GROUND Noun Frequency	1.57	0.28
13	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.08	0.20
14	PC10: GROUND Noun Picture Size	0.95	0.17
15	PC1: Topological Nominal Frequency	0.74	0.13
16	PC6: Entropy	0.60	0.11
17	PC5: RE	0.58	0.11
18	PC4: Phrase Frequency	0.45	0.08

B.3 Experiment 2: Late Locative sentences

B.3.1 Preposition

B.3.1.1 Probability

Table B.37: Model summary. Fixation Probability for the preposition for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	0.037	0.091	0.403	0.687
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	25.775	29.000	218.485	< 0.001
s(Item,bs="re")	38.129	156.000	50.464	0.004
s(Trial)	1.000	1.000	8.273	0.004
s(PC1: Topological Nominal Frequency)	1.000	1.001	5.329	0.021

Table B.38: GBM variable importance. Fixation Probability for the preposition for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (2.78%).

rank	predictor	rel. infl	abs. infl.
1	Participant	37.24	2.43
2	Item	28.23	1.84
3	Trial	15.52	1.01
4	PC10: GROUND Noun Picture Size	3.07	0.20
5	PC11: Topological Nominal Length	2.75	0.18
6	PC4: Phrase Frequency	2.37	0.15
7	PC2: GROUND Noun Frequency	2.02	0.13
8	PC6: Entropy	1.99	0.13
9	PC1: Topological Nominal Frequency	1.55	0.10
10	PC5: RE	1.31	0.09
11	PC8: GROUND Noun Length	1.19	0.08
12	PC3: GROUND Noun Visual Complexity	1.08	0.07
13	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.86	0.06
14	PC7: Topological Nominal Visual Complexity	0.82	0.05

B.3.1.2 Position

Table B.39: Model summary. Fixation Position for initial fixations on the preposition for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	18.056	0.518	34.867	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	22.608	29.000	3.294	< 0.001
s(Item,bs="re")	28.461	156.000	0.228	0.028
s(Previous Fixation Duration)	1.000	1.000	6.430	0.011
s(Partial Saccade Length)	2.924	2.994	146.494	< 0.001
s(PC7: Top. Nominal Visual Complexity)	2.312	2.634	3.789	0.016

Table B.40: GBM variable importance. Fixation Position for initial fixations on the preposition for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (15.43%).

rank	predictor	rel. infl	abs. infl.
1	Partial Saccade Length	43.74	9.43
2	Item	18.94	4.09
3	Participant	8.77	1.89
4	Previous Fixation Duration	6.39	1.38
5	Trial	5.37	1.16
6	Y Position	5.17	1.12
7	PC11: Topological Nominal Length	1.95	0.42
8	PC3: GROUND Noun Visual Complexity	1.62	0.35
9	PC8: GROUND Noun Length	1.45	0.31
10	PC2: GROUND Noun Frequency	1.29	0.28
11	PC7: Topological Nominal Visual Complexity	1.05	0.23
12	PC5: RE	1.00	0.22
13	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.96	0.21
14	PC1: Topological Nominal Frequency	0.91	0.20
15	PC10: GROUND Noun Picture Size	0.58	0.13
16	PC6: Entropy	0.56	0.12
17	PC4: Phrase Frequency	0.23	0.05

B.3.1.3 Duration

Table B.41: Model summary. Fixation Duration for initial fixations on the preposition for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.291	0.033	159.505	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	27.177	29.000	16.768	< 0.001
s(Trial)	2.150	2.538	3.347	0.021
s(Saccade Length)	2.587	2.884	9.120	< 0.001

Table B.42: GBM variable importance. Fixation Duration for initial fixations on the preposition for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (12.31%).

rank	predictor	rel. infl	abs. infl.
1	Participant	36.27	8.52
2	Item	16.49	3.87
3	Saccade Length	10.67	2.51
4	Trial	7.93	1.86
5	Previous Fixation Duration	6.60	1.55
6	Y Position	6.03	1.42
7	X Position	3.35	0.79
8	PC4: Phrase Frequency	2.30	0.54
9	PC3: GROUND Noun Visual Complexity	2.20	0.52
10	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.45	0.34
11	PC7: Topological Nominal Visual Complexity	1.37	0.32
12	PC8: GROUND Noun Length	1.11	0.26
13	PC6: Entropy	1.05	0.25
14	PC5: RE	1.05	0.25
15	PC1: Topological Nominal Frequency	0.59	0.14
16	PC2: GROUND Noun Frequency	0.57	0.13
17	PC10: GROUND Noun Picture Size	0.53	0.12
18	PC11: Topological Nominal Length	0.45	0.10

B.3.2 GROUND noun

B.3.2.1 Probability

Table B.43: Model summary. Fixation Probability for the GROUND noun for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	2.525	0.227	11.146	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	26.583	29.000	436.720	< 0.001
s(Item,bs="re")	27.291	145.000	34.913	0.015
s(Trial)	1.000	1.000	19.308	< 0.001
s(PC2: GROUND Noun Frequency)	2.539	2.784	29.083	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	40.268	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	4.632	0.031
s(PC8: GROUND Noun Length)	1.431	1.661	14.382	< 0.001
s(PC10: GROUND Noun Picture Size)	1.000	1.000	4.346	0.037

Table B.44: GBM variable importance. Fixation Probability for the GROUND noun for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (15.06%).

rank	predictor	rel. infl	abs. infl.
1	Participant	47.55	10.00
2	Item	32.56	6.85
3	Trial	12.02	2.53
4	PC3: GROUND Noun Visual Complexity	1.99	0.42
5	PC1: Topological Nominal Frequency	1.29	0.27
6	PC11: Topological Nominal Length	1.27	0.27
7	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.91	0.19
8	PC8: GROUND Noun Length	0.54	0.11
9	PC5: RE	0.50	0.11
10	PC7: Topological Nominal Visual Complexity	0.42	0.09
11	PC4: Phrase Frequency	0.39	0.08
12	PC6: Entropy	0.21	0.04
13	PC10: GROUND Noun Picture Size	0.19	0.04
14	PC2: GROUND Noun Frequency	0.15	0.03

B.3.2.2 Position

Table B.45: Model summary. Fixation Position for initial fixations on the GROUND noun for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	26.944	1.010	26.668	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	27.624	29.000	19.272	< 0.001
s(Item,bs="re")	71.491	139.000	1.105	< 0.001
s(Trial)	1.942	2.327	8.355	< 0.001
s(Previous Fixation Duration)	2.575	2.876	6.743	< 0.001
s(Partial Saccade Length)	2.941	2.997	513.167	< 0.001
s(PC1: Topological Nominal Frequency)	1.495	1.640	34.732	< 0.001
s(PC3: GROUND Noun Visual Complexity)	2.664	2.780	7.044	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	4.929	0.026
s(PC8: GROUND Noun Length)	1.749	1.919	39.421	< 0.001
s(PC9: GROUND Noun Freq. (SUBTLEX))	1.000	1.000	4.332	0.037
s(PC10: GROUND Noun Picture Size)	1.000	1.000	8.743	0.003
s(PC11: Topological Nominal Length)	1.122	1.173	4.203	0.041

Table B.46: GBM variable importance. Fixation Position for initial fixations on the GROUND noun for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (33.22%).

rank	predictor	rel. infl	abs. infl.
1	Partial Saccade Length	45.78	18.11
2	Item	22.02	8.71
3	Participant	18.68	7.39
4	Previous Fixation Duration	3.76	1.49
5	Trial	2.26	0.89
6	PC8: GROUND Noun Length	1.67	0.66
7	Y Position	1.26	0.50
8	PC1: Topological Nominal Frequency	1.19	0.47
9	PC3: GROUND Noun Visual Complexity	0.67	0.26
10	PC7: Topological Nominal Visual Complexity	0.64	0.25
11	PC5: RE	0.43	0.17
12	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.38	0.15
13	PC11: Topological Nominal Length	0.37	0.15
14	PC10: GROUND Noun Picture Size	0.32	0.12
15	PC4: Phrase Frequency	0.28	0.11
16	PC6: Entropy	0.17	0.07
17	PC2: GROUND Noun Frequency	0.13	0.05

B.3.2.3 Duration

Table B.47: Model summary. Fixation Duration for initial fixations on the GROUND noun for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.396	0.039	138.387	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	28.232	29.000	37.045	< 0.001
s(Item,bs="re")	48.584	155.000	0.465	< 0.001
s(Trial)	1.883	2.267	6.749	0.001
s(X Position)	2.221	2.601	3.423	0.016
s(Y Position)	1.629	1.998	2.941	0.050
s(Saccade Length)	2.192	2.595	5.859	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	13.389	< 0.001
s(PC4: Phrase Frequency)	1.000	1.000	5.526	0.019

Table B.48: GBM variable importance. Fixation Duration for initial fixations on the GROUND noun for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (18.66%).

rank	predictor	rel. infl	abs. infl.
1	Participant	55.56	14.24
2	Item	10.08	2.58
3	Saccade Length	5.94	1.52
4	X Position	5.80	1.49
5	Previous Fixation Duration	5.04	1.29
6	Y Position	3.67	0.94
7	Trial	2.97	0.76
8	PC1: Topological Nominal Frequency	2.67	0.68
9	PC8: GROUND Noun Length	1.23	0.31
10	PC3: GROUND Noun Visual Complexity	1.18	0.30
11	PC9: GROUND Noun Frequency (SUBTLEX-CH)	1.08	0.28
12	PC10: GROUND Noun Picture Size	1.05	0.27
13	PC5: RE	0.80	0.20
14	PC4: Phrase Frequency	0.77	0.20
15	PC2: GROUND Noun Frequency	0.69	0.18
16	PC6: Entropy	0.65	0.17
17	PC11: Topological Nominal Length	0.48	0.12
18	PC7: Topological Nominal Visual Complexity	0.36	0.09

B.3.3 Nominal

B.3.3.1 Probability

Table B.49: Model summary. Fixation Probability for the topological nominal for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	z-value	p-value
(Intercept)	-0.712	0.194	-3.669	< 0.001
smooth terms	edf	ref. df	Chi sq.	p-value
s(Participant,bs="re")	27.807	29.000	530.564	< 0.001
s(Item,bs="re")	37.664	141.000	52.150	0.002
s(PC1: Topological Nominal Frequency)	1.000	1.000	593.604	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	10.142	0.001
s(PC4: Phrase Frequency)	1.000	1.000	78.369	< 0.001
s(PC5: RE)	1.000	1.000	7.805	0.005
s(PC7: Top. Nominal Visual Complexity)	1.000	1.000	7.893	0.005
s(PC8: GROUND Noun Length)	1.000	1.000	5.884	0.015
s(PC9: GROUND Noun Freq. (SUBTLEX))	1.000	1.000	6.856	0.009
s(PC11: Topological Nominal Length)	1.000	1.000	88.589	< 0.001

Table B.50: GBM variable importance. Fixation Probability for the topological nominal for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (32.87%).

rank	predictor	rel. infl	abs. infl.
1	Item	60.44	22.75
2	Participant	30.36	11.43
3	Trial	4.82	1.81
4	PC1: Topological Nominal Frequency	0.76	0.29
5	PC4: Phrase Frequency	0.72	0.27
6	PC8: GROUND Noun Length	0.55	0.21
7	PC10: GROUND Noun Picture Size	0.42	0.16
8	PC3: GROUND Noun Visual Complexity	0.38	0.14
9	PC7: Topological Nominal Visual Complexity	0.35	0.13
10	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.34	0.13
11	PC5: RE	0.28	0.10
12	PC11: Topological Nominal Length	0.23	0.09
13	PC6: Entropy	0.21	0.08
14	PC2: GROUND Noun Frequency	0.15	0.05

B.3.3.2 Position

Table B.51: Model summary. Fixation Position for initial fixations on the topological nominal for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	19.167	0.703	27.251	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	22.855	29.000	4.384	< 0.001
s(Item,bs="re")	44.009	144.000	0.541	< 0.001
s(Trial)	2.666	2.910	3.436	0.015
s(Y Position)	2.201	2.590	3.296	0.023
s(Partial Saccade Length)	2.860	2.982	148.313	< 0.001
s(PC1: Topological Nominal Frequency)	2.382	2.617	41.544	< 0.001
s(PC8: GROUND Noun Length)	1.000	1.000	15.066	< 0.001
s(PC11: Topological Nominal Length)	1.000	1.000	47.388	< 0.001

Table B.52: GBM variable importance. Fixation Position for initial fixations on the topological nominal for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (30.55%).

rank	predictor	rel. infl	abs. infl.
1	Item	34.16	11.88
2	Partial Saccade Length	33.84	11.77
3	Participant	12.36	4.30
4	Y Position	4.35	1.51
5	Previous Fixation Duration	3.61	1.26
6	Trial	2.21	0.77
7	PC11: Topological Nominal Length	2.10	0.73
8	PC1: Topological Nominal Frequency	0.97	0.34
9	PC4: Phrase Frequency	0.95	0.33
10	PC5: RE	0.94	0.33
11	PC7: Topological Nominal Visual Complexity	0.92	0.32
12	PC8: GROUND Noun Length	0.85	0.29
13	PC6: Entropy	0.69	0.24
14	PC2: GROUND Noun Frequency	0.68	0.24
15	PC3: GROUND Noun Visual Complexity	0.56	0.19
16	PC10: GROUND Noun Picture Size	0.53	0.19
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.27	0.09

B.3.3.3 Duration

Table B.53: Model summary. Fixation Duration for initial fixations on the topological nominal for Late Locative sentences in Experiment 2.

parametric coefficients	Est.	S.E.	t-value	p-value
(Intercept)	5.216	0.034	151.690	< 0.001
smooth terms	edf	ref. df	F-value	p-value
s(Participant,bs="re")	25.538	29.000	8.774	< 0.001
s(Trial)	1.000	1.000	15.543	< 0.001
s(X Position)	1.000	1.000	44.566	< 0.001
s(Y Position)	2.333	2.711	3.336	0.032
s(Previous Fixation Duration)	1.377	1.659	5.110	0.018
s(Saccade Length)	2.114	2.525	2.998	0.026
s(PC1: Topological Nominal Frequency)	1.000	1.000	66.984	< 0.001
s(PC3: GROUND Noun Visual Complexity)	1.000	1.000	7.712	0.005
s(PC4: Phrase Frequency)	1.000	1.000	8.113	0.004
s(PC11: Topological Nominal Length)	1.000	1.000	5.459	0.020

Table B.54: GBM variable importance. Fixation Duration for initial fixations on the topological nominal for Late Locative sentences in Experiment 2. Relative influences (rel. infl.) sum to 100. Corrected influences (abs. infl.) sum to the percentage of variance explained by the model (11.68%).

rank	predictor	rel. infl	abs. infl.
1	Participant	27.30	5.31
2	Item	17.15	3.34
3	X Position	11.78	2.29
4	Saccade Length	8.08	1.57
5	Previous Fixation Duration	7.77	1.51
6	Trial	7.52	1.46
7	PC1: Topological Nominal Frequency	6.09	1.18
8	Y Position	5.08	0.99
9	PC4: Phrase Frequency	1.47	0.29
10	PC11: Topological Nominal Length	1.35	0.26
11	PC10: GROUND Noun Picture Size	1.32	0.26
12	PC5: RE	1.15	0.22
13	PC7: Topological Nominal Visual Complexity	0.91	0.18
14	PC6: Entropy	0.81	0.16
15	PC3: GROUND Noun Visual Complexity	0.75	0.15
16	PC8: GROUND Noun Length	0.71	0.14
17	PC9: GROUND Noun Frequency (SUBTLEX-CH)	0.45	0.09
18	PC2: GROUND Noun Frequency	0.30	0.06

References

Academia Sinica. (1998). *Academia Sinica Balanced Corpus (CD-ROM, version 3).* Taipei, Taiwan: Academia Sinica.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge, U.K.: Cambridge University Press.

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., & Schreuder, R. (1999). War and peace: morphemes and full forms in a non-interactive activation parallel dual route model. *Brain and Language*, *68*, 27-32.

Baayen, R. H., & Schreuder, R. (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences)*, *358*, 1-13.

Baayen, R. H., Vasishth, S., Bates, D., & Kliegl, R. (2016). The cave of shadows. addressing the human factor with generalized additive mixed models. *Manuscript.*

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. I., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods*, *34*(3), 424–434.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statisistical Society*, *26*, 211–246.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *JEP:HPP*, *42*(3), 441–458.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6). Available from `e10729 .doi:10.1371/journal.pone.0010729`

Center for Chinese Linguistics. (2006). *Chinese Corpus of the Center of Chinese Linguistics at Peking University.* Available from `http://ccl.pku.edu.cn: 8080/ccl_corpus/index.jsp?dir=xiandai`

Chang, P. C., Galley, M., & Manning, C. (2008). Optimizing Chinese word segmentation for machine translation performance. *Proceedings of the Third Workshop on Statistical Machine Translation*, 224–232.

Chang, Y. N., Hsu, C. H., Chen, C. L., & Lee, C. Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods*, *48*(1), 112–122.

Chen, H. C., Song, H., Lau, W. Y., Wong, K. F. E., & Tang, S. L. (2003). Developmental characteristics of eye movements in reading Chinese. In C. McBride-Chang & H. C. C. Chen (Eds.), *Reading development in Chinese children* (pp. 157–169). Westport, CT: Praeger.

Chen, H. C., & Tzeng, O. J. L. (1992). *Language Processing in Chinese*. Amsterdam: North-Holland.

Chen, H. C., Vaid, J., & Wu, J. T. (2009). Homophone density and phonological frequency in Chinese word recognition. *Language and Cognitive Processes*, *24*(7-8), 967–982.

Chen, J. Y., & Dell, G. S. (2006). Word-form encoding in Chinese speech production. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (Vol. 1, pp. 165–174). New York: Cambridge University Press.

Chen, M. J., & Weekes, B. S. (2004). Effects of semantic radicals on Chinese character categorization and character decision. *Chinese Journal of Psychology*, *46*, 179–195.

Chen, M. J., Weekes, B. S., Peng, D. L., & Lei, Q. (2006). Effects of semantic radical consistency and combinability on Chinese character processing. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (Vol. 1, pp. 175–186). New York: Cambridge University Press.

Chen, T., He, T., & Benesty, M. (2015). xgboost: eXtreme Gradient Boosting [Computer software manual]. Available from `https://github.com/dmlc/xgboost` (R package version 0.4-0)

Chen, W. F., Chao, C., P, Chang, Y. N., & Hsu, C. H. (2016). Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain and Language*, *157-158*.

Cheng, C. M. (1981). Perception of Chinese characters. *Acta Psychologica Taiwanica*, *23*, 137–153.

Chinese Academy of Social Sciences. (2012). *Contemporary Chinese Dictionary* (6th ed.). China: The Commercial Press.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–258.

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica*, *32*, 133–143.

DeFrancis, J. (1984). *The Chinese Language: Facts and Fantasy.* Honolulu: University of Hawaii Press.

Dimitropoulou, M., Duñabeitia, J., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behaviour: the case of Greek. *Frontiers in Psychology*, *1:218*, 1–12.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 2, pp. 1212–1248). Berlin, Boston: De Gruyter Mouton.

Fan, K. Y., Gao, J. Y., & Ao, X. P. (1984). Pronunciation principles of the Chinese character and alphabetic writing scripts. *Chinese Character Reform*, *3*, 23–27.

Fang, S. P., Horng, R. Y., & Tzeng, O. J. L. (1986). Consistency effects in the Chinese character and pseudo-character naming tasks. In H. S. R. Kao & R. Hoosain (Eds.), *Linguistics, Psychology, and the Chinese Language* (pp. 11–21). Hong Kong: Center of Asian Studies, University of Hong Kong.

Feldman, L. B., & Siok, W. W. T. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*, 778-781.

Feldman, L. B., & Siok, W. W. T. (1999a). Semantic radicals contribute to the visual identification of Chinese characters. *Journal of Memory and Language*, *40*, 559-576.

Feldman, L. B., & Siok, W. W. T. (1999b). Semantic radicals in phonetic compounds: Implications for visual character recognition in Chinese. In J. Wang, A. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis.* Hillsdale, NJ: Erlbaum.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., et al. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488-496.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks, CA: Sage. Available from `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, *59*, 127–136.

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.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674–691.

Graff, D., & Chen, K. (2003). *Chinese Gigaword LDC2003T09.* Philadelphia: Linguistic Data Consortium.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A Mosaic of Corpus Linguistics: Selected Approaches* (pp. 269–291). Frankfurt am Main: Peter Lang.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1(3)*, 297-318.

Hayes, B. (2009). *Introductory Phonology.* UK: Wiley.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Hebb, D. O. (1949). *The Organization of Behavior.* New York: Wiley & Sons.

Heister, J., Würzner, K., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB - eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*, *62*, 10–20.

Hendrix, P. (2016). *Experimental explorations of a discrimination learning approach to language processing.* Unpublished doctoral dissertation, Eberhard Karl's Universität, Tübingen.

Hendrix, P., Bolger, P., & Baayen, R. H. (2016). Distinct ERP signatures of word frequency, phrase frequency and prototypicality in speech production. *Journal of Experimental Psychology: Language, Memory and Cognition*, *in press*. Available from `10.1037/a0040332`

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*, 353-367.

Honorof, D. N., & Feldman, L. (2006). The Chinese character in psycholinguistic research: form, structure and the reader. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (Vol. 1, pp. 195–217). New York: Cambridge University Press.

Hoosain, R. (1991). *Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese*. Hillsdale, NJ: Erlbaum.

Howie, J. M. (1974). On the domain of tone in Mandarin: Some acoustical evidence. *Phonetica*, *30*, 129–148.

Hsieh, S. K. (2006). *Hanzi, concept and computation: A preliminary survey of Chinese characters as a knowledge resource in NLP (doctoral dissertation)*. Eberhard Karls Universität Tübingen.

Hsu, C. H., Tsai, J. L., Lee, C. Y., & Tzeng, O. J. L. (2009). Orthographic combinability and phonological consistency effects in reading Chinese phonograms: an event-related potential study. *Brain and Language*, *108*, 56–66.

Huang, H. W., Lee, C. Y., Tsai, J. L., Lee, C. L., Hung, D. L., & Tzeng, O. J. L. (2006). Orthographic neighborhood effects in reading Chinese two-character words. *Neuroreport*, *17*(10), 1061–1065.

Hue, C. (1992). Recognition processes in character naming. In E. Chen & O. Tzeng (Eds.), *Language Processing in Chinese* (pp. 93–107). Amsterdam: North-Holland.

Inhoff, A. W. (1999). Use of prelexical and lexical information during Chinese sentence reading: evidence from eye-movement studies. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 223-238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Inhoff, A. W., & Liu, W. (1997). The perceptual span during the reading of chinese text. In H. C. Wang (Ed.), *The Cognitive Processing of Chinese and Related Asian Languages*. Hong Kong: The Chinese University Press.

Inhoff, A. W., & Liu, W. (1998). The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 20–34.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 229–254). Amsterdam: John Benjamins.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*(174).

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Kuo, W. J., Yeh, T. C., Lee, C. Y., Wu, Y., Chou, C. C., Ho, L. T., et al. (2003). Frequency effects of Chinese character processing in the brain: an event-related fMRI study. *NeuroImage*, *18*(3), 720–730.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83-97.

Lee, C. Y., Hsu, C. H., Chang, Y. N., Chen, W. F., & Chao, P. C. (2015). The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguist norm. *Language and Linguistics*, *16*(4), 535–554.

Lee, C. Y., Tsai, J. L., Kuo, W. J., Yeh, T. C., Wu, Y. T., Ho, L. T., et al. (2004). Neuronal correlates of consistency and frequency effects in Chinese character naming: an event-related fMRI study. *NeuroImage*, *23*(4), 1235–1245.

Leong, C. K., Cheng, P. W., & Mulcahy, R. (1987). Automatic processing of morphemic orthography by mature readers. *Language and Speech*, *30*(2), 181–196.

Li, P., Tan, L. H., Bates, E., & Tzeng, O. J. L. (2006). *The Handbook of East Asian Psycholinguistics* (Vol. 1). New York: Cambridge University Press.

Liu, I. M. (1988). Context effects on word/character naming: Alphabetic versus logographic languages. In I. M. Liu, H. C. Chen, & M. J. Chen (Eds.), *Cognitive Aspects of the Chinese Language* (p. 81-92). Hong Kong: Asian Research Service.

Liu, I. M. (1999). Character and word recognition in Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 173-187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Liu, W., Inhoff, W., A., Ye, Y., & Wu, C. (2002). Use of parafoveally visible characters during the reading of Chinese sentences. *Journal of Experimental Psychology. Human Perception and Performance*, *28*, 1213–1227.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods*, *39*(2), 192–198.

Liu, Y., Zhang, L. J., & Shu, H. (2006). The role of initial phoneme on naming latency [in Chinese]. *Psychological Science*, *29*, 64–67.

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*(2), 471–483.

McEnery, A. M., & Xiao, R. Z. (2003). *The Lancaster Corpus of Mandarin Chinese*. Paris, France / Oxford, UK: European Language Resources Association / Oxford Text Archive.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*. Available from `http://arxiv.org/abs/1301.3781`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Milin, P., Filipović Durdević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. (2009). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in Grammar: Form and Acquisition* (pp. 214–252). Oxford: Oxford University Press.

Ministry of Education of the People's Republic of China. (1988). 现代汉语通用字表 *[List of Commonly Used Characters in Modern Chinese]*.

Ministry of Education of the People's Republic of China. (2013). 通用规范汉字表 *[Table of General Standard Chinese Characters]*.

Miwa, K., Libben, G., Dijkstra, T., & Baayen, R. H. (2014). The time-course of lexical activation in Japanese morphographic word recognition: Evidence for a character-driven processing model. *Quarterly Journal of Experimental Psychology*, *67*(1), 79–113.

National Language Commission of China (国家语委). (1997). *Standard Stroke Order of Commonly-Used Characters of Modern Chinese (*现代汉语通用字笔顺规范*)*. Beijing: Language & Culture Press.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods*, *36*, 516-524.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: Lexique. *L'Année Psychologique*, *101*, 447–462.

Parkvall, M. (2007). Världens 100 största sprak [the world's largest 100 languages]. In *Nationalencyklopedin.* Malmö: NE Nationalencyklopedin AB.

Peng, D. L., Liu, Y., & Wang, C. M. (1999). How is access representation organized? the relation of polymorphemic words and their components in Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 65-89). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Perfetti, C. A., & Liu, Y. (2006). Reading Chinese characters: Orthography, phonology, meaning, and the lexical constituency model. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (Vol. 1, pp. 225–236). New York: Cambridge University Press.

Perfetti, C. A., & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(1), 101–118.

Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese character identification. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 115-134). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pham, H. (2014). *Visual processing of Vietnamese compound words: A multivariate analysis of using corpus linguistic and psycholinguistic paradigms (doctoral dissertation).* University of Alberta.

Pham, H., & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, *30*.

Ramscar, M., Dye, M., & McCauley, S. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760–793.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

Rayner, K., Li, X., Juhasz, B. J., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review*, *12*, 1089–1093.

Rayner, K., & Pollatsek, A. (1989). *The Psychology of Reading.* Eaglewood Cliffs, NJ: Prentice-Hall.

Rayner, K., Well, A., Pollatsek, A., & Bertera, J. (1982). The availability of useful information to the right of fixation in reading. *Perception and Psychophysics*, *31*, 537–550.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research [Computer software manual]. Evanston, Illinois. Available from `http://CRAN.R-project.org/package=psych` (R package version 1.5.8)

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (p. 131-154). Hillsdale, New Jersey: Lawrence Erlbaum.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, *19*, 1–30.

Shaoul, C., Sun, C. C., & Ma, J. Q. (2016). The Simplified Chinese Corpus of Webpages (SCCoW). *Manuscript.*

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus.* Bologna: Gedit.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323–348.

Sun, C. (2006). *Chinese: A Linguistic Introduction.* United Kingdom: Cambridge University Press.

Sun, H. L., Huang, J. P., Sun, D. J., Li, D. J., & Xing, H. B. (n.d.). Introduction to language corpus system of modern Chinese study. In *Paper collection for the fifth world Chinese teaching symposium.* Beijing: Peking University Publisher.

Sze, W., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese Lexicon Project: a repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*(1), 263–273.

Taft, M. (2006). Processing of characters by native Chinese readers. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (Vol. 1, pp. 237–249). New York: Cambridge University Press.

Taft, M., Huang, J., & Zhu, X. (1994). The influence of character frequency on word recognition responses in Chinese. In H. W. Chang, J. T. Huang, C. W. Hue, & O. J. L. Tzeng (Eds.), *Advances in the study of Chinese language processing* (Vol. 1, p. 59-73). Taipei: Department of Psychology, National Taiwan University.

Taft, M., Liu, Y., & Zhu, X. (1999). Morphemic processing in reading Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (pp. 91–114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Taft, M., & Zhu, X. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*(3), 761–775.

Torgo, L. (2010). *Data Mining Using R: Learning With Case Studies.* New York: CRC Press. Available from `http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR`

Tsai, J. L., Kliegl, R., & Yan, M. (2012). Parafoveal semantic information extraction in traditional Chinese reading. *Acta Psychologica*, *141*(1), 17–23.

Tsai, J. L., Lee, C. H., Lin, Y. C., Tzeng, O. J. L., & Hung, D. L. (2006). Neighborhood size effects of Chinese words in lexical decision and reading. *Language and Linguistics*, *7*(3), 659–675.

Tsai, J. L., Lee, C. Y., Tzeng, O. J. L., Hung, D. L., & Yen, N. S. (2004). Use of phonological codes from Chinese characters: Evidence from processing of parafoveal preview when reading sentences. *Brain and Language*, *91*, 235–244.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190.

Vitevich, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*(4), 735–747.

Wang, H. C., Pomplun, M., Chen, M., Ko, H., & Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability. *The Quarterly Journal of Experimental Psychology*, *63*(7), 1374–1386.

Wang, J., Inhoff, A. W., & Chen, H. C. (1999). *Reading Chinese Script: A Cognitive Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W., Ning, N., & Zhang, J. X. (2012). The nature of the homophone density effects: an ERP study with Chinese spoken monosyllabic homophones. *Neuroscience Letters*, *516*(1), 67-71.

Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organising maps in R: the kohonen package. *Journal of Statistical Software*, *21*(5).

Wikipedia. (2016). *Help: IPA for Mandarin — Wikipedia, the free encyclopedia*. Available from `http://en.wikipedia.org/wiki/Help:IPA_for_Mandarin` ([Online; accessed 25-January-2016])

Wood, S. (2006). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analysis does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Xiao, H. (2010-2015). 汉语拼音标注工具. `http://www.cncorpus.org/`.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61–83.

Yan, G., Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology*, *97*, 259–268.

Yan, M., Richter, E. M., Shu, H., & Kliegl, R. (2009). Chinese readers extract semantic information from parafoveal words during reading. *Psychonomic Bulletin and Review*, *16*, 561–566.

Yan, M., Zhou, W., Shu, H., & Kliegl, R. (2012). Lexical and sub-lexical semantic preview benefits in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1069–1075.

Yang, H. M., & McConkie, G. W. (1999). Reading Chinese: some basic eye-movement characteristics. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 207-222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Yang, J., Wang, S., Xu, Y., & Rayner, K. (2009). Do Chinese readers obtain preview benefit from word n + 2? Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1192–1204.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992–1003.

Yeo, I., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*, 954–959.

Yip, P. C. (2000). *The Chinese Lexicon: A Comprehensive Survey*. New York: Routledge.

Yu, A. C. L. (2010). Tonal effects on perceived vowel duration. *Laboratory Phonology*, *10*.

Zhang, B. Y., & Peng, D. L. (1992). Decomposed storage in the Chinese lexicon. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language Processing in Chinese* (pp. 131–149). Amsterdam: North-Holland.

Zheng, X. (1983). Is it easy to learn Chinese characters? *Educational Research*, *4*, 56–63.

Zhou, X., & Marslen-Wilson, W. (1999). Sublexical processing in reading Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese Script: A Cognitive Analysis* (p. 37-63). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ziegler, J., Tan, L. H., Perry, C., & Montant, M. (2000). Phonology matters: the phonological frequency effect in written Chinese. *Psychological Science*, *11*(3), 234–238.

Zipf, G. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

Summary

This thesis presents a new large-scale lexical database for simplified Chinese and investigates lexical processing at the character level, the word level, and the phrase level through the lexical variables contained in this database. The database is called the Chinese Lexical Database (henceforth CLD) and contains lexical information about 30,645 one-character words and two-character words (4,710 one-character words and 25,935 two-character words). For each of the words in the CLD, 164 lexical variables are provided. These lexical variables are either categorical (23 predictors) or numerical (141 predictors) in nature. The CLD is publicly available and can be downloaded and searched at `http://www.chineselexicaldatabase.com`.[1]

The lexical variables in the CLD include frequency measures at the word level and the character level, as well as below the character level (i.e., semantic radical frequency, phonetic radical frequency, visual component frequency). Furthermore, the CLD contains visual complexity measures at different grain sizes (i.e., visual component counts, stroke counts, pixel counts, picture sizes), as well as orthographic neighbourhood density measures. Phonological properties of words and characters are available in the CLD as well and include phonological frequency measures, phonological complexity measures and phonological neighbourhood density measures. In addition, the CLD provides various measures of the orthography-to-phonology and phonology-to-orthography consistency, both at the character level (homograph and homophone type and token counts and frequencies) and below the character level (phonetic radical regularity and consistency measures). Finally, the CLD contains a number of information-theoretic measures that tap into the combinatorial properties of characters (e.g., entropy, trigram entropy, relative entropy, mutual information).

[1]The website `http://www.chineselexicaldatabase.com` is password-protected until this dissertation is published. The password is `75090246`.

The data in the CLD provide lexical information for simplified Chinese that allows psycholinguistic researchers to efficiently establish which lexical characteristics influence response variables in individual experiments and to compare the results of different experiments in a consistent manner. This thesis presents two examples of the use of the lexical information in the CLD to gain insight into lexical processing at the character level, the word level, and the phrase level. First, I investigated response times, pronunciation durations, and eye fixation durations in a single-participant word naming study. Second, I gauged lexical processing of locative phrases through an analysis of the eye movement patterns in multi-participant sentence reading study. Locative phrases are the semantic equivalent of prepositional phrases in English, and consist of the preposition 在, a GROUND noun and a topological nominal. The data for both analyses were analyzed using state-of-the-art statistical methods: gradient boosting machines and generalized additive mixed-effect models.

The two experiments showed the expected effects of frequency and visual complexity, both at the character level and at the word level. Visual complexity influenced response variables at multiple grain sizes. Furthermore, I found effects of the phonological frequency, complexity, and neighbourhood density on the duration of fixations in the word naming task. Lexical characteristics of the phonetic radical and the semantic radical also influenced measures of lexical processing in the word naming task, although their effects were more subtle than the effects of predictors at the character level and the word level. The most striking effects, however, were related to the combinatorial properties of characters and words. Both in the word naming task and in the phrase reading task, I observed robust effects of entropy and relative entropy. At the word level, high values of entropy and relative entropy resulted in additional processing costs, as indicated by more, longer, and more leftward eye fixations. At the character level, the effects of entropy and relative entropy were reversed, with shorter naming latencies and reduced eye fixation durations for characters with high entropies and high relative entropies. The reversal of the entropy effect at the character level may be a result of the increased orthography-to-phonology consistency for characters that combine with many other characters. Furthermore, I observed facilitatory phrase frequency effects on the eye fixation patterns in the phrase reading task. To my knowledge, the effects of entropy, relative entropy and phrase frequency are the first effects of these predictors reported for simplified Chinese. These effects suggest that the language processing system is sensitive to information-theoretic properties of the language.

In addition to predictor effects, the analyses of the eye movement data for both experiments revealed two important facts about lexical processing. First, lexical items are not processed in a strictly serial fashion. Instead, we found consistent evidence for the joint processing of multiple lexical items, both at and above the word level. Joint processing of lexical items was possible due to parafoveal preview, which allowed for the uptake of information from words or characters that were not in the center of visual attention. Second, a driving force behind lexical processing is the dynamic allocation of resources to information-rich areas. Eye fixation durations are longer when the word or character that is fixated on provides more information and shorter when neighbouring words or characters provide more information. Furthermore, words are fixated on more often when they provide more information and less often when upcoming words provide more information. Finally, fixations are less far into words that provide more information and further into the current word when the upcoming words provide more information. Lexical processing in Chinese therefore is a highly dynamic process, in which upcoming information is pre-processed through parafoveal preview and in which information-theoretic properties of the input guide language users towards maximally efficient eye movement patterns.