The role of information theory for compound words in Mandarin Chinese and English

Peter Hendrix
University of Tübingen

Ching Chu Sun
University of Tübingen

We investigate the role of information-theoretic measures for compound word reading in two languages: Mandarin Chinese and English. For each language, we report the results of two analyses: a time-to-event analysis using piece-wise additive mixed models (PAMMs) and a causal inference analysis with causal additive models (CAMs). We use the PAMM analyses to gain insight into the temporal profile of the effects of information-theoretic measures in the word naming task. For both English and Mandarin Chinese, we observed early effects of the entropy of both constituents, as well as temporally widespread effects of point-wise mutual information (PMI). The CAM analyses provide further insight into the relations between lexical-distributional variables. The image that emerges from the CAM analyses is that the information-theoretic measures entropy and PMI are embedded in a carefully balanced system in which lexical-distributional properties that lead to processing difficulties are offset by lexical-distributional properties that guarantee successful communication. The information-theoretic measures have a central position in this system, and are causally influenced not only by frequency, but also by the effects of other lower-level lexical-distributional variables such as visual complexity, and phonology-to-orthography consistency.

Keywords: information theory, compounds, entropy, mutual information, Mandarin Chinese, PAMM, CAM

Introduction

In 1948, Claude Shannon proposed information theory in his seminal paper "A Mathematical Theory of Communication" (Shannon, 1948). Information theory provides a mathematical description of the concept of information in human and non-human communication. In the last decade-and-a-half, nearly sixty years after Shannon's first publication on the subject, interpretations of behavioral data in language processing from the perspective of information theory have started to emerge. These interpretations typically focus on the effects of measures derived from information theory on a behavioral measure of language processing, such as the reaction time in a lexical decision experiment or the acoustic duration of a word. The information-theoretic measures either encode the amount of uncertainty in the signal (e.g., entropy), or, conversely, the extent to which uncertainty is reduced by the signal (e.g., association measures, conditional probability). The extent to which the signal reduces uncertainty is referred to as information.

Entropy measures the uncertainty in the signal, quantified as the average number of bits required to encode a signal. The number of bits of a signal is defined as the log of the inverse of its probability. Entropy is the sum over the number of bits of the potential signals, weighted for the probability of these messages:

$$H = \sum_{i=1}^{n} P(i) \cdot log_2 \frac{1}{P(i)} \tag{1}$$

where P_i is the probability of message i.

The more similar the probabilities of the potential signals, the greater the uncertainty about the signal, and the higher the entropy. As an example, consider the probability distribution of compounds in which the left constituent is "star". The first column of Table 1 provides the frequency (per million words) in SUBTLEX-UK (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) for the 5 noun-noun compounds in the English Lexicon Project (henceforth ELP Balota et al., 2007) for which the modifier is "star": "starboard", "starlight", "stardust", "starfish", and "stargazer".

The second column of Table 1 converts the frequencies into probabilities ($P_i = \frac{f_i}{\sum_{i=1}^{n} f_i}$). Next, we calculate

Corresponding author: Ching Chu Sun (ching-chu.sun@uni-tuebingen.de)

Table 1
Entropy for compounds with the left constituent "star".

compound	f_i	$P(i)$	$\log_2 \frac{1}{P(i)}$	$P(i) \cdot \log_2 \frac{1}{P(i)}$
starboard	6.55	0.675	0.567	0.383
starlight	1.59	0.164	2.608	0.428
stardust	0.78	0.080	3.644	0.292
starfish	0.75	0.077	3.699	0.285
stargazer	0.04	0.004	7.966	0.032
sum	9.71	1.000	-	1.419

the binary logarithm of the inverse probabilities, which represents the number of bits required to encode each of the compounds. The binary logarithm of the inverse probabilities is presented in the third column of Table 1. To calculate the average number of bits required to encode a compound that starts with "star", the number of bits required to encode each compound is weighted for the probability of the compound. This reflects the fact that the need to encode "starboard" arises more often than the need to encode "starfish" or "stargazer". Summing across the probability-weighted number of bits required to encode each compound in the fourth column of Table 1 yields the entropy over the probabilities for the noun-noun compounds in which the left constituent is "star", which is 1.419.

Bien, Levelt, and Baayen (2005) explored the role of entropy in compound processing. The authors measured onset latencies for productions of Dutch compound words in a position-response association task. Higher values for the entropy of both the left constituent (as in Table 1) and the right constituent corresponded to shorter naming latencies. Response times thus were shorter for compounds that consist of constituents that occur in a higher number of compounds with more similar frequencies.

The relevance of entropy for compound processing was confirmed by Kuperman, Pluymaekers, Ernestus, and Baayen (2007), who investigated the effect of entropy on acoustic durations of interfixes in Dutch compounds. The authors analysed the acoustic duration of linking morphemes in Dutch compounds. In the Dutch compound "ballenbak" ("ball pit"), for instance, the constituents ("ball") and "bak" ("pit", literally "bin" or "container") are linked through the interfix "-en". The greater the entropy of the right constituent of a compound, they found, the longer the acoustic duration of the interfix in that compound.

Recently, Schmidtke, Kuperman, Gagné, and Spalding (2016) looked into the role of entropy on compound processing in English. Rather than defining entropy over form-level probabilities, Schmidtke et al. (2016) calculated the entropy over the probabilities of structural relationships between the modifier and the head of a compound (e.g.; "MADE OF", "CAUSED BY", ...). The higher this entropy, the authors found, the longer the response time in a lexical decision experiment. Greater uncertainty about the relational structure of a compound thus corresponds to additional processing costs.

The effects of entropy reported by Schmidtke et al. (2016) are opposite in nature to the effects of entropy reported by Bien et al. (2005). A number of potential explanations for this discrepancy exist. First, the diverging results could be due to the fact that the entropy was calculated over different structures in both studies. The entropy measures in (Bien et al., 2005) were defined at the form level, whereas Schmidtke et al. (2016) calculated entropy over the conceptual relations in a compound. Furthermore, the former study calculated entropies at the constituent level, whereas the latter defined entropy for compounds as a whole. Both studies may therefore tap into different aspects of the combinatorial properties of constituents in compounds.

Second, the different effects of entropy in both studies could be task-related. Bien et al. (2005) investigated the role of entropy on language processing in a task that gauges aspects of language production. By contrast, Schmidtke et al. (2016) adopted the lexical decision paradigm, which taps into processes that play a role in language comprehension. The qualitatively different effects of entropy in both studies could therefore arise due to the different demands the tasks in these studies impose on the language processing system. A third possibility is that the way in which information-theoretic measures manifest themselves in behavioral data depends on the distributional structure of the language under investigation. Differences may exist between the distributional space for compound words in English and Dutch, which could lead differences in how compounds are processed in both languages.

As noted above, entropy gauges the amount of uncertainty in the signal. Uncertainty is inversely proportional to information (which can, indeed, be conceptualised as uncertainty reduction). Conditional probabilities and association measures quantify the amount of information in the signal. Kuperman, Bertram, and Baayen (2008) argued that the conditional probability of the right constituent of a compound given the left constituent (i.e., the probability of "stardust" given the fact that the left constituent of the compound is "star" ($P(stardust \mid star) = 0.080$) plays an important role in compound processing. The relevance of association measures for compound processing has, to our knowledge, not previously been examined. The association measure mutual information, which compares the frequency of two linguistic elements occurring together to the frequency both elements occurring in isolation, however, has been shown to influence acoustic durations at both the word level (Pluymaekers, Ernestus, & Baayen, 2005) and the segment level (Kuperman, Ernestus, & Baayen, 2008).

The effects of information-theoretic measures on compound processing indicate that the language processing system is sensitive to the combinatorial properties of the constituents in compound words. As noted above, however, questions remain about the exact manner in which paradigmatic relations in compound words influence that way in which we perceive and produce these morphologically complex words. Do the effects of entropy for Dutch compound in a response-association task reported by Bien et al. (2005) generalize to the word naming task? Is the influence of information-theoretic measures on compound processing limited to the effect of entropy or do other measures, such as mutual information have an effect on behavioral measures of compound processing as well? Are the effects of information-theoretic measures a general property of linguistic processing across languages, or do the effects of these measures depend on the structure and the distributional properties of a language?

Below, we seek to gain more insight into these issues through an investigation of the effects of a number of information-theoretic measures on compound processing in the reading aloud task in English and Mandarin Chinese. In English, compound words are a relatively rare phenomenon. Of all $79,686$ words in the ELP, $3,501$ words are compound words, for a type percentage of 4.39%. The token percentage of compounds is even lower. The contribution of compound words to the total token count of words that are in the ELP in SUBTLEX-UK is no more than 1.29%, which indicates that compound words tend to be relatively low frequency words. It should be noted that the ELP does not include compounds in which the constituents are separated by a hyphen or a space. The true rate of occurrence of compound words is therefore higher than the estimates provided here. Nonetheless, it is clear that compound words constitute a minority among lexical items in English.

Mandarin Chinese is a tonal language. The basic phonological unit is the syllable. A 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). In writing, syllables are expressed through characters. Recent estimates indicate that 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). Honorof and Feldman (2006) estimated that about a third of the word tokens in Mandarin Chinese are mono-syllabic and consist of a single character. An overwhelming majority of the remaining two thirds of word tokens in Chinese are di-syllabic words that consist of two characters.

The majority of two-character words in Mandarin Chinese are compound words, although affixation is a productive process for two-character words in Mandarin Chinese as well (Myers, 2006). Morphemes tend to appear in multiple compound words, in which they form various functional relations with other constituent morphemes (Zhou & Marslen-Wilson, 1995). As is the case in English, the right constituent is the head of the compound, whereas the left constituent is the modifier. The word for "classroom" in Mandarin Chinese, for instance, is "教室". The left constituent "教" means teach, whereas the right constituent "室" means room (example adopted from X. Chen, Hao, Geva, Zhu, & Shu, 2009).

The role of compound words, however, is much more prominent in Mandarin Chinese than in English. Compounding is the most productive word formation process in Mandarin Chinese. Furthermore, compounds are much more frequent in Mandarin Chinese than in English. Of all $48,644$ words in the Chinese Lexical Database (henceforth CLD; C. C. Sun, Hendrix, Ma, & Baayen, 2018), $34,233$ words are di-syllabic two-character words, for a type percentage of 70.37%. As noted above, most of these two-character words are compounds words, although exceptions do exist. Together, the $34,233$ two-character words in the CLD account for 36.85% of the word tokens in the corpora underlying the frequency counts in the CLD. Due to the central position of compound words in Mandarin Chinese, speakers of this language have much more experience with compound words as compared to speakers of English. One of the questions we seek to answer in the current work is to what extent differences in exposure to compound words influence the way in which speakers process these words.

In what follows, we provide two analyses of the word naming latencies for compound words in Mandarin Chinese and English. First, we report the results of an analysis using a statistical technique that is embedded in the tradition of time-to-event analysis: the piece-wise exponential additive mixed model (henceforth PAMM Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, Hartl, Day, & Küchenhoff, 2018). Like standard regression models, PAMMs estimate the effect of one or more predictors on a response variable. Whereas the response variable in a standard regression model of naming latencies is the end-point of processing, however, PAMMs estimate the probability of an instantaneous response as it develops over time. As a result, PAMMs provide information about the temporal dynamics of predictor effects that is not available through standard regression models. The aim of the PAMM analysis is to establish whether or not information-theoretic measures influence compound processing in the word naming task in English and Mandarin Chinese, and to determine to what extent the qualitative nature and the temporal development of the effects of these predictors differs between both languages. The PAMM analysis of the data thus focuses on the influence of information-theoretic properties of compounds on language processing.

The second analysis uses the causal additive model (henceforth CAM; Peters, Mooij, Janzing, & Schölkopf, 2014) to gain more insight into lexical-distributional space in Mandarin Chinese and English. The causal additive model falls under the statistical umbrella of causal inference models, and seeks to establish causal connections between lexical-distributional variables. The CAM analysis focuses on the position of information-theoretic properties of compound words in distributional space in English and Mandarin Chinese. How do information-theoretic measures relate to other lexical-distributional properties of compound words, such as visual complexity or frequency? Are there systematic relations between lexical-distributional properties of words? Do these relationships exist across languages, or are they language-specific? The CAM analysis of the word naming data in Mandarin Chinese and English promises to provide more insight into the answers to these questions.

Chinese

Methods

Participants. Three participants took part in the experiment: two females and one male. All three participants were native speakers of Mandarin Chinese and had normal or corrected-to-normal vision. The average age of the participants was 34.

Materials. We created a list of one-character and two-character 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` (Peterson, 2005). We excluded two-character words that are a repetition of the same character, proper nouns, and Japanese Kanji (Hanzi). This resulted in a stimulus set of 4,710 one-character words and 25,935 two-character words, for a total of 30,645 words. The experiment consisted of 31 experiment lists. The first 30 lists consisted of 1,000 trials, whereas the last list consisted of 645 trials. Here, we focus on the naming latencies for the 25,935 two-character words.

Design. The response variable in our analysis of the naming latencies is the temporal onset of the pronunciation of each word. Across the three participants, the total number of trials for two-character words is $3 \times 25,935 = 77,805$. Prior to analysis, we removed incorrect responses from the data. This led to the exclusion of 4,420 data points (5.68%). After outlier removal, the data set thus consisted of 73,385 response times. We investigated the effect of a number of predictors on the naming latencies. The lexical information for the predictors under investigation was extracted from the Chinese Lexical Database.

First, we added the initial phoneme (*Initial Phoneme*) of the word as a control variable in our analyses. We furthermore included three frequency measures: the frequency of the word as a whole, the frequency of the first character, and the frequency of the second character. The frequency counts in the CLD are based on combined frequencies in SUBTLEX-CH (Cai & Brysbaert, 2010) and the Leiden Weibo Corpus (henceforth LWC; Van Esch (2012)). SUBTLEX-CH is 33.5 million word corpus of film subtitles, whereas the LWC is a corpus of messages posted on the social medium Sina Weibo that consists of 101.4 million words. Prior to analysis, a logarithmic transform was applied to all frequency accounts to remove a rightward skew from the frequency distributions. Henceforth, we therefore refer to the three frequency measures as *(log) Frequency*, *(log) C1 Frequency*, and *(log) C2 Frequency*.

In addition, two measures of visual complexity were entered into the model. These measures of visual complexity encode the number of strokes in the first and second character. 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. To reduce asymmetry in the stroke counts distributions, we applied square root transforms to the stroke counts prior to analysis. We thus refer to these measures of visual complexity as *(sqrt) C1 Strokes* and *(sqrt) C2 Strokes*.

The effect of *(sqrt) C2 Strokes*, however, did not reach significance and is hence omitted from the description of the results of our analysis.

Our primary interest for the current purposes, however, is in information-theoretic measures. Two types of information-theoretic measures were entered into the model: entropy and mutual information. For a given two-character word, *C1 Entropy* entropy is defined as the entropy over the probabilities of all two-character that share the first character with the target word. Analogously, *C2 Entropy* is the entropy over the probabilities of all two-character words that share the second character with the target word.

The word 苹果 ("apple"), for instance, is the only two-character word in the CLD that starts with the character 苹. As a result, there is no uncertainty about the second character when we know that the first character of a two-character word is 苹. Hence, *C1 Entropy* for the word 苹果 ("apple") is 0. By contrast, there are 32 two-character words in the CLD that end with the character 果, including 9 words with a frequency greater than 10 per million words. Substantial uncertainty about the identity of the first character thus remains when we know that the second character of a two-character word is "果". Consequently, *C1 Backward Entropy* is higher than *C1 Entropy* the word 苹果 ("apple"): 1.89. Prior to analysis, we applied square root transformations to the entropy measures to reduce asymmetry in their distributions. Henceforth, we therefore refer to these measures as *(sqrt) C1 Entropy* and *(sqrt) C2 Entropy*.

Pointwise mutual information (henceforth PMI) is a measure of the strength of the association between both characters in a two-character word. PMI compares the observed frequency of a two-character words to its expected frequency (c.f., Gries, 2010). The expected frequency is defined as:

$$\frac{\sum_{j=1}^{O} f_i \sum_{k=1}^{P} f_j}{\sum_{i=1}^{N} f_i} \tag{2}$$

where N is the set of two-character words in the lexicon, O and P are the sets of two-character words that share the first and second character with the target word, and f_i is the frequency of word i.

For example, the observed frequency per million for the word 苹果 ("apple") is 107.66. The PMI measure used here is position-specific. The expected frequency of the word 苹果 ("apple") therefore depends on the summed frequency of all two-character words (334227.90), the summed frequency of all two-character words that start with the character 苹 (107.66; only the word 苹果 ("apple") itself), and the summed frequency of all two-character words that end with the character 果 (1788.38; 32 words). The expected frequency for the word 苹果 ("apple") thus is $\frac{107.66 \times 1788.38}{334227.90} = 0.57$.

PMI is defined as the logged ratio of the observed and expected frequency of a word:

$$\log_2\left(\frac{\text{observed frequency}}{\text{expected frequency}}\right) \tag{3}$$

.

As such, the value of the predictor *PMI* for the word 苹果 ("apple") is $\log_2\left(\frac{107.66}{0.57}\right) = 7.56$. No transformation was applied to *PMI* prior to analysis.

Readers may ask themselves why we opted to use entropy measures rather than measures of the morphological family size of a compound's constituents. Morphological family size is the type count of the number of compounds that share either the left (*C1 Family Size*) or the right (*C2 Family Size*) constituent with the target word. Similar to the entropy measures described above, morphological family size is a measure of the combinatorial properties of constituents in compounds. Several studies revealed that greater morphological families correspond to shorter response latencies in a variety of tasks, including lexical decision and word naming (see, e.g., Juhasz & Berkowitz, 2011; De Jong, Schreuder, & Baayen, 2000). Consistent with these findings, facilitatory effects of morphological family size have been observed for Mandarin Chinese as well, both in word naming (Y. Liu, Shu, & Li, 2007) and in lexical decision (Tsai, Lee, Lin, Tzeng, & Hung, 2006). Effects of the summed frequency of all members of a family, the morphological family frequency, have been reported for compound processing as well, both in alphabetical languages (De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002) and in Mandarin Chinese (Huang et al., 2006).

The raw correlations of family size and family frequency with the observed naming latencies (*(log) C1 Family Size*: $r = -0.208$, *(log) C2 Family Size*: $r = -0.133$, *(log) C1 Family Frequency*: $r = -0.225$, *(log) C2 Family Frequency*: $r = -0.150$) are higher than the raw correlations of entropy with the observed naming latencies (*(sqrt) C1 Entropy*: -0.161, *(sqrt) C2 Entropy*: $r = -0.108$). This is, however, due to the increased correlations of these measures with the corresponding character frequencies (*(log) C1 Family Size*: $r = 0.869$, *(log) C2 Family Size*: $r = 0.859$, *(log) C1 Family Frequency*: 0.961, *(log) C2 Family Frequency*: $r = 0.962$), as compared to the entropy measures (*(sqrt) C1 Entropy*: $r = 0.583$, *(sqrt) C2 Entropy*: $r = 0.613$). Indeed, once character frequencies are taken into account, the effects of the entropy measures (*(sqrt) C1 Entropy*: $t = -14.286$, *(sqrt) C2 Entropy*: $t = -10.081$) in a linear regression model are more prominent than the effects

Table 2
Distributional statistics for the predictors (log) Frequency, (log) C1 Frequency, (log) C2 Frequency, (sqrt) C1 Strokes, (sqrt) C1 Entropy, (sqrt) C2 Entropy, and PMI. For each predictor, we provide the original range and the adjusted range after outlier removal, as well as the mean, the median and standard deviation after outlier removal.

predictor	range	adj. range	mean	median	sd
(log) Frequency	-4.21-8.38	-4.21-7.14	0.18	0.12	1.96
(log) C1 Frequency	-3.92-10.39	-3.92-8.62	5.40	5.58	1.74
(log) C2 Frequency	-4.62-10.39	-4.62-8.51	5.48	5.71	1.68
(sqrt) C1 Strokes	1.00-5.00	1.00-4.24	2.83	2.83	0.56
(sqrt) C1 Entropy	0.00-2.31	0.00-2.31	1.53	1.62	0.45
(sqrt) C2 Entropy	0.00-2.51	0.00-2.51	1.58	1.64	0.48
PMI	-8.51-24.43	-7.10-18.09	5.46	5.45	4.14

of family size (*(log) C1 Family Size*: $t = -11.469$, *(log) C2 Family Size*: $t = -9.059$) and family frequency (*(log) C1 Family Frequency*: $t = -4.459$, *(log) C2 Family Frequency*: $t = -4.250$). For the current data, the entropy measures described above thus provided more explanatory power than do measures of the family size or family frequency of a compound's constituents.

Prior to the PAMM analysis we removed predictor outliers further than 3 standard deviations from the predictor mean from the data. This resulted in the exclusion of 0.05% of the data (38 observations) for *(log) Frequency*, 1.70% of the data ($1,247$ observations) for *(log) C1 Frequency*, 1.79% of the data ($1,311$ observations) for *(log) C2 Frequency*, 0.56% of the data (411 observations) for *(sqrt) C1 Strokes*, and 0.30% of the data (218 observations) for *PMI*. No outliers further than 3 standard deviations from the predictor mean were present for *(sqrt) C1 Entropy* and *(sqrt) C2 Entropy*. Table 2 presents the distributional statistics for the predictors that entered the analysis. Provided are original ranges and ranges after outlier removal, as well as means, median, and standard deviations of all predictors. The exclusion of predictor outliers reduced the data set by $3,124$ data points (4.26% of the data). The data set for the PAMM analysis thus consisted of $70,261$ data points.

Procedure. The experiment took place in a sound-proof booth. Participants were instructed to respond as fast as possible, while retaining accuracy. Prior to each trial a fixation mark was shown in the center of the screen. Next, a word was presented in the center of the screen in black SimHei 80 point font. 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. A 10 minute break was inserted halfway through each experimental session. Naming latencies were extracted from the recorded speech signal through custom computer code on the basis of volume thresholds. The performance of this code was inspected on a trial-by-trial basis and corrected manually where necessary.

PAMM analysis

For the analysis of the word naming latencies in Mandarin Chinese we use a piece-wise exponential additive mixed model (PAMM Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018). The PAMM is a statistical technique for time-to-event analysis (which is referred to as survival analysis as well). Time-to-event analyses model the time until an event of interest occurs. The event of interest in the word naming task is the onset of the pronunciation, which corresponds to the naming latency. The primary advantage of time-to-event analysis over traditional analysis techniques for response time data is that predictor effects may be modelled as a function of time, even though the response time itself measures the end-point of processing only. Unlike traditional analysis techniques, time-to-event analysis thus provides insight into the temporal dynamics of lexical processing. Below, we introduce the main concepts behind the PAMM. For a more comprehensive introduction to the use of PAMMs in response time studies, we refer the interested reader to Hendrix (2019). A formal introduction to the PAMM and its statistical properties is provided in Bender, Groll, and Scheipl (2018). Bender and Scheipl (2018) provide useful practical examples of the application of PAMMs for data sets with different types of predictors.

Whereas traditional regression analyses model the response time itself, the PAMM models the probability of a response as it develops over time. More precisely, the PAMM models the hazard function $\lambda(t)$, which describes the instantaneous probability of a response at time t, provided that no response was registered prior to time t:

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T \leq t + d \mid T \geq t)}{dt} \tag{4}$$

where T is the response time.

As noted above, we collected word naming data for three participants. Whereas the standard deviation of the response time distribution was similar across the three participants (range: 93.02 - 107.91), the median response time differed substantially between the participants. The median response time for the fastest participant was 488.30 ms, whereas the median response time for the slowest participant was 640.09 ms. When fit to the raw response times, the hazard function would thus be dominated by responses of the fastest participant for low values of t and by responses of the slowest participant for high values of t. To prevent this, we shifted the response times distributions for each participant. The median response time of the shifted response time distributions for each participant is identical to the median response time in the full data set: 514.51 ms.

The (log) hazard function for the shifted word naming data as modelled through a PAMM is presented in Figure 1. The instantaneous probability of a response (i.e., the onset of the pronunciation) rapidly increases between 370 and 500 ms after stimulus onset. After that, the (log) hazard rate remains relatively stable throughout the remainder of the response time window. The shape of the hazard function observed here is typical for response time distributions in psycholinguistic experiments (cf. Hendrix, 2018, 2019).

We limited the response time window from 370 ms to 1,000 ms after stimulus onset. This response time window contains 99.16% of the responses. It is important to note, however, that response times shorter than 370 ms or longer than 1,000 ms remain part of the analysis. The exact response times for these words are unknown to the model. The model does know, however, that these stimuli were responded to before 370 ms after stimulus onset and after 1,000 ms after stimulus onset, respectively. Valuable information about the stimuli that were particularly easy or hard to respond to thus remains available to the model.

The overall hazard function for the data is referred to as the baseline hazard. The baseline hazard itself often

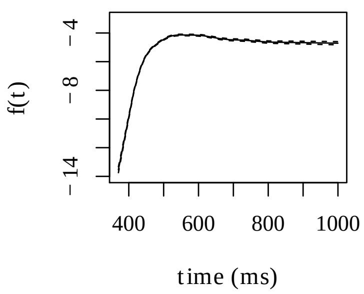

Figure 1. Log-transformed instantaneous hazard function ($f(t)$) with point-wise confidence intervals for Mandarin Chinese.

is of little interest. Here, too, our primary interest is in how the values of predictors modulate the shape of the hazard function. The PAMM is an extremely flexible framework that allows for the modeling of such modulations. As a semi-parametric extension of the generalized additive mixed model (GAMM; S. Wood, 2011; S. N. Wood, 2017), it inherits the ability of the GAMM to model non-linear effects in multiple dimensions. The PAMM is therefore able to model non-linear predictor effects that evolve in a non-linear fashion as a function of time.

The PAMM models the hazard function in a piece-wise fashion for each of a number of intervals in the time dimension. Following Hendrix (2019), we split the response time window into 51 intervals that are evenly spread out across the quantiles of the response time distribution. The (log) hazard function $\lambda(t|x_i)$ for all time points t in the interval $j := (\kappa_{j-1}, \kappa_j]$ given the predictor values x_i for stimulus i is defined as:

$$\log(\lambda(t|x_i)) = \log\lambda_0(t_j) + \sum_{k=1}^{p} f_k(x_{i,k}, t_j) + b_\ell, \quad \forall t \in (\kappa_{j-1}, \kappa_j]. \tag{5}$$

where $\lambda_0(t_j)$ is the baseline hazard for time interval j, $f_k(x_{i,k}, t_j)$ are smooth functions for predictor $k \in 1, \ldots, p$ for each time point t in the interval j, and b_ℓ are random intercepts associated with group $\ell \in 1, \ldots, L$ to which stimulus i belongs.

The smooth functions $f_k(x_{i,k}, t_j)$ allow the PAMM to model non-linear predictor effects, whereas the piece-wise nature of the fitting process enables the PAMM to model non-linearities in the time dimension. In a practical sense, non-linear predictor effects that evolve in a non-linear fashion over time can be modelled through tensor product interactions (cf. S. Wood, 2011) between time and the predictor in question. Despite the fact such effects are modelled through (two-dimensional) smooth functions, however, the estimates of the PAMM remain piece-wise constant in the time dimension.

We fitted a PAMM for the word naming latencies with the `mgcv` package for R (S. Wood, 2011; S. N. Wood, 2017). The PAMM models the instantaneous probability of a response as a function of time and the predictors *Initial Phoneme*, *(log) Frequency*, *(log) C1 Frequency*, *(log) C2 Frequency*, *(sqrt) C1 Strokes*, *(sqrt) C1 Entropy*, *(sqrt) C2 Entropy*, and *PMI*. The baseline hazard in Figure 1 and time-constant predictor effects were estimated by smooth terms. Time-varying predictor effects were estimated by tensor product interactions, as modelled through `ti()` terms (see S. N. Wood, 2017, for more details).

No restrictions were imposed on the main effect smooth for time that models the baseline hazard. In the interest of interpretability of the results, however, we limited predictor smooths as well as time by predictor tensor product terms to fourth order non-linearities. We did not add a random effect of participant to the model, because the participant-specific hazard functions were nearly identical after the participant-specific shifting of the response time distributions mentioned above.

Medium-strength correlations exist between the predictors under investigation, as indicated by a relatively high condition number of $\kappa = 40.581$ (Belsley, Kuh, & Welsch, 1980). We therefore fit separate PAMMs with a main effect smooth for time, a main effect smooth of the predictor and a tensor product interaction between time and the predictor for each of the predictor under investigation. The predictor effects in these PAMMs were highly similar to the predictor effects in the full model, which indicates that the influence of collinearity on the results reported below was modest.

Table 3
Results of the piece-wise exponential additive mixed model (PAMM) fit to the naming latencies. For parametric terms, β estimates, standard errors of the β estimates and p-values are shown. For smooth terms, estimated degrees of freedom, χ^2 values and p-values are provided.

parametric terms	β	**S.E.**	p
Intercept	-5.145	0.044	< 0.001
smooth terms	**edf**	χ^2	p
time	8.989	4259.878	< 0.001
InitialPhoneme	28.484	2668.634	< 0.001
(log) Frequency	2.826	1420.041	< 0.001
time by *(log) Frequency*	5.172	151.713	< 0.001
(log) C1 Frequency	2.938	363.610	< 0.001
time by *(log) C1 Freq.*	6.840	491.981	< 0.001
(log) C2 Frequency	2.663	50.689	< 0.001
time by *(log) C2 Freq.*	5.378	46.355	< 0.001
(sqrt) C1 Strokes	2.548	122.055	< 0.001
time by *(sqrt) C1 Strokes*	5.072	103.522	< 0.001
C1 Entropy	2.658	177.937	< 0.001
time by *C1 Entropy*	7.364	122.079	< 0.001
C2 Entropy	2.822	142.892	< 0.001
time by *C2 Entropy*	2.294	38.307	< 0.001
PMI	1.010	328.546	< 0.001
time by *PMI*	3.349	42.242	< 0.001

PAMM results

The results for the PAMM fit to the word naming data in Mandarin Chinese are presented in Table 3. The baseline hazard of the model differs significantly from zero, as indicated by a significant model intercept ($\beta =$ -116.850, $p < 0.000$), as well as a significant main effect smooth of *time* ($\chi^2 = 4259.878$, $p < 0.000$). The baseline hazard of the PAMM was presented in Figure 1 above. As noted above, the baseline hazard increases rapidly from 370 to 500 ms after stimulus onset, after which it stabilises until the end of the analysis window. We furthermore observed a significant effect of the control variable *Initial Phoneme* ($\chi^2 = 2668.634$, $p < 0.000$).

The baseline hazard is modified by the frequency of a word. The main effect of *(log) Frequency* is significant ($\chi^2 = 1420.041$, $p < 0.000$), as is the *time* by *(log) Frequency* interaction ($\chi^2 = 151.713$, $p < 0.000$). The partial main effect of *(log) Frequency* is presented in Figure 2. The y-axis in Figure 2 shows adjustments to the baseline hazard as a function of *(log) Frequency.* This adjustment is negative for low frequency words and positive for high frequency words. The main effect of *(log) Frequency* thus suggests that the instantaneous probability of a response is higher for high frequency words as compared to low frequency words throughout the analysis window.

The main effect of *(log) Frequency*, however, is modulated by the interaction between *time* and *(log) Frequency.* Figure 3 visualizes the partial interaction between *time* and *(log) Frequency.* The time after stimulus onset is on the x-axis of Figure 3, whereas the (log of the) frequency of the word is on the y-axis. The color coding represents the adjustment to the baseline hazard, with warmer colors representing positive adjustments and colder colors representing negative adjustments. Faded areas correspond to points in time where the interaction between *time* and *(log) Frequency* did not reach significance (i.e., points in time where 0 was in the confidence interval for all predictor values). White areas indicates the absence of responses at a certain point in time. From 370 to 500 ms after stimulus onset, hazard rates are significantly higher for high frequency words. During later stages of the response window, however, the nature of the effect reverses, with a lower instantaneous probability of a response for high frequency words as compared to low frequency words.

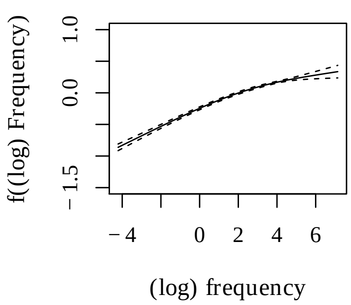

Figure 2. Partial main effect of *(log) Frequency* for Mandarin Chinese.

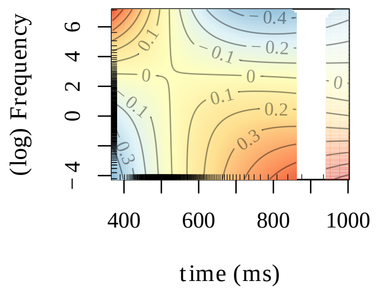

Figure 3. Partial interaction of *time* and *(log) Frequency* for Mandarin Chinese. Warmer colors indicate higher (log) hazard rates.

The full adjustment to the baseline hazard as a function of word frequency is obtained by adding up the partial main effect of *(log) Frequency* and the partial interaction between *time* and *(log) Frequency*. The full effect of *(log) Frequency* is presented in Figure 4. The effect of *(log) Frequency* is most prominent during the early stages of the response window, with increased hazard rates for high frequency words. The size of the effect of *(log) Frequency* decreases as a function of time, and the effect is no longer significant at 941 ms after stimulus onset. The probability of an instantaneous response, therefore, is higher for high frequency words as compared to low frequency words, particularly during the early stages of the response window. The effect of word frequency reported here fits well with the results of previous studies, in which the facilitatory effect of word frequency was solidly established in both lexical decision (Lee, Hsu, Chang, Chen, & Chao, 2015; Sze, Rickard Liow, & Yap, 2014; Zhang & Peng, 1992; Peng, Liu, & Wang, 1999) and word naming (Seidenberg, 1985; Y. Liu et al., 2007; I. M. Liu, 1999).

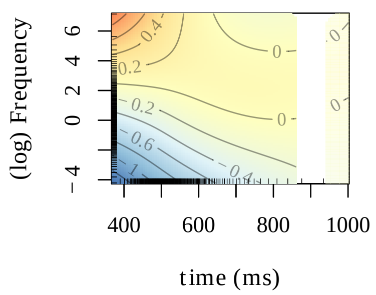

Figure 4. Effect of *(log) Frequency* for Mandarin Chinese. Warmer colors indicate higher (log) hazard rates.

Not only the frequency of words as whole influences behavioral measures of lexical processing. Effects of the frequency of the characters within words have been reported as well. Zhang and Peng (1992), Taft, Huang, and Zhu (1994), and Peng et al. (1999) all reported facilitatory effects of character frequency in lexical decision. Yan, Tian, Bai, and Rayner (2006) observed a character frequency effect on eye fixation durations on two-character words. (Kuo et al., 2003) and Lee et al. (2004) found character frequency effects in fMRI studies. Studies that failed to observe character frequency effects, however, exist as well. Janssen, Bi, and Caramazza (2008) did not find constituent frequency effects in a picture naming task in both English and Mandarin Chinese, whereas T. M. Chen and Chen (2006) reported the absence of constituent frequency effects in a response-association task.

For the word naming data under investigation here, we observed a significant main effect of *(log) C1 Frequency* ($\chi^2 = 363.610$, $p < 0.000$), as well as a significant *time* by *(log) C1 Frequency* interaction ($\chi^2 = 491.981$, $p < 0.000$). The effect of *C1 Frequency* is presented in Figure 5. As was the case *(log) Word Frequency*, the effect of *(log) C1 Frequency* is most prominent in the early stages of the response window, although it is significant throughout the response window (i.e., from 370 to 1000 ms after stimulus onset). Consistent with the facilitatory effects of character frequency reported in the literature, hazard rates are higher for words with more frequent first characters.

The PAMM analysis revealed an effect of the frequency of the second character as well, with a significant main effect of *(log) C2 Frequency* ($\chi^2 = 50.689$, $p < 0.000$) and a significant interaction of *time* and *(log) C2 Frequency* ($\chi^2 = 46.355$, $p < 0.000$). Hazard rates are significantly higher for words with frequent second characters from 370 to 764 ms after stimulus onset. The effect of the frequency of the second character thus is more transient

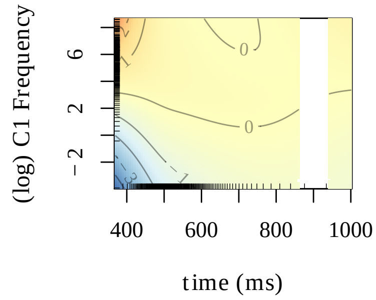

Figure 5. Effect of *(log) C1 Frequency* for Mandarin Chinese.

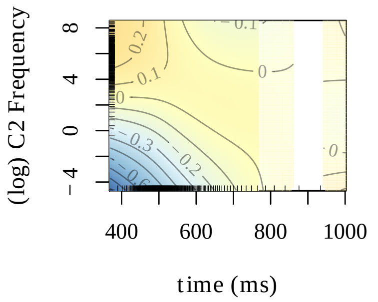

Figure 6. Effect of *(log) C2 Frequency* for Mandarin Chinese.

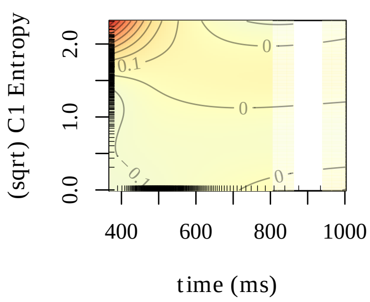

Figure 8. Effect of *(sqrt) C1 Entropy* for Mandarin Chinese.

than the effect of the frequency of the first character. As can be seen in Figure 6, the effect size of the effect of *(log) C2 Frequency* is considerably smaller than the effect size of *(log) C1 Frequency*. This reflects the increased prominence of the first character as compared to the second character as a result of the left-to-right uptake of information in reading.

The visual complexity of a word influences the instantaneous probability of a response as well. As noted above, however, the effect of visual complexity is limited to an effect of the number of strokes in the first character. As can be seen in Table 3, both the main effect of *(sqrt) C1 Strokes* (χ^2 = 122.055, $p < 0.000$) and the interaction between *time* by *(sqrt) C1 Strokes* ($\chi^2 = 103.522$, $p < 0.000$) reached significance. The adjustment to the baseline hazard as a function of *(sqrt) C1 Strokes* is presented in Figure 7. Hazard rates are higher for words with fewer strokes from 370 to 569 ms after stimulus onset. When a word was not responded to at 569 ms after stimulus onset, the visual complexity of the first character thus no longer has a significant influence on the decision making process. The effect of

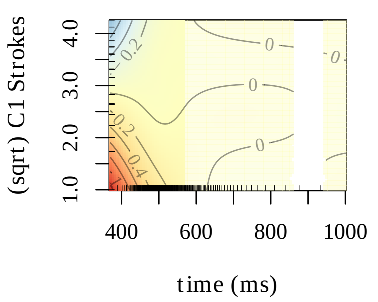

Figure 7. Effect of *(sqrt) C1 Strokes* for Mandarin Chinese.

(sqrt) C1 Strokes reported here is in line with the inhibitory effects of stroke counts in both lexical decision (Lee et al., 2015) and word naming (Y. Liu et al., 2007; Leong, Cheng, & Mulcahy, 1987) that were reported in previous studies.

The focus of the current study is on the effects of the information-theoretic measures entropy and PMI. The PAMM analysis revealed significant main effects and significant interactions with time for the entropy of both the first (*(sqrt) C1 Entropy*: χ^2 = 177.937, $p < 0.000$, *time* by *(sqrt) C1 Entropy*: $\chi^2 = 122.079$, $p < 0.000$) and the second character (*(sqrt) C2 Entropy*: χ^2 = 142.892, $p < 0.000$, *time* by *(sqrt) C2 Entropy*: χ^2 = 38.307, p < 0.000).

The effects of *(sqrt) C1 Entropy* and *C2 Entropy* are presented in Figure 8 and Figure 9. As can be seen in Figure 8, the effect of *(sqrt) C1 Entropy* is significant from 370 to 804 ms after stimulus onset. The effect is most prominent during the early stages of the response time window, with a higher instantaneous probability of a response for high values of *(sqrt) C1 Entropy*. Figure 9 indicates that the effect of *(sqrt) C2 Entropy* is qualitative similar, with higher hazard rates for high values of *(sqrt) C2 Entropy*. The effect of *(sqrt) C2 Entropy*, however, reaches significance from 370 to 588 ms after stimulus onset only. Furthermore, the effect size of the effect of *(sqrt) C2 Entropy* is smaller than the effect size of the effect of *(sqrt) C1 Entropy*. As was the case for the effects of character frequency, the effect of entropy thus is more pronounced for the first character than for the second character.

Despite the structural differences between Dutch and Mandarin Chinese, the effects of entropy observed here are similar to the effects of entropy reported by Bien et al. (2005) for compound processing in Dutch in a position-response association task. Consistent with the current findings, the authors of this study found that higher constituent-level entropies resulted in shorter

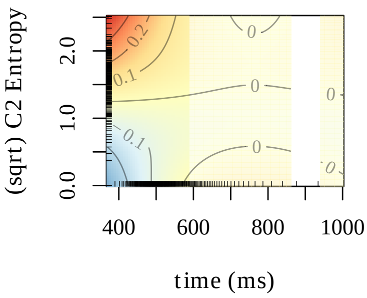

Figure 9. Effect of *(sqrt) C2 Entropy* for Mandarin Chinese.

naming latencies. Both studies thus revealed a processing advantage for compounds that consist of constituents that occur in a higher number of compounds with more similar frequencies. This suggests that the additional experience for characters that occur in a large number of two-character compound words leads to a processing advantage.

The second type of information-theoretic measure under investigation is *PMI*: a measure of the association between the characters in a compound. As was the case for the effects of the entropy of both constituent, both the main effect of *PMI* ($\chi^2 = 328.546$, $p < 0.000$) and the *time* by *PMI* interaction ($\chi^2 = 42.242$, $p < 0.000$) were highly significant. The effect of *PMI* is presented in Figure 10. The instantaneous probability of a response is higher for lower values of *PMI*. The effect of *PMI* is less transient than the effects of entropy and remains significant throughout the analysis window.

The effect of *PMI* indicates that a stronger association between the characters in a two-character words results in additional processing difficulties. At first glance, this may seem surprising. One might expect a processing advantage for two-character compound words that consist of characters that nearly always occur together. Given one character in these words, little uncertainty remains about the other character. On the other hand, however, readers have more experience with characters that occur in a larger number of two-character words. This increased experience might make words that contains such character easier to process. Furthermore, there may be lexical-distributional differences between two-character words with high and low values of *PMI* that help better understand the inhibitory effect of *PMI*. In the next section, we explore the lexical-distributional space for Mandarin Chinese compound words in more detail through a CAM analysis of the word naming data.

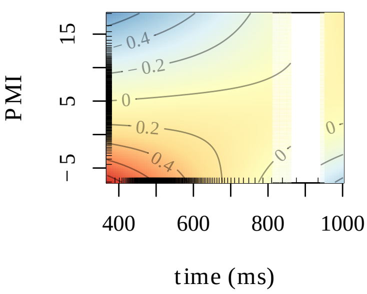

Figure 10. Effect of *PMI* for Mandarin Chinese.

CAM analysis

We investigated lexical-distributional space for compound words in Mandarin Chinese using a causal inference model. Causal inference models attempt to infer causal relations between predictors. The causal structure in a causal inference model is often visually represented through a graph. A graph consists a series of vertexes connected by edges. The edges in a causal graph are directed. A directed edge from vertex i to vertex j indicates a causal relationship between the predictors associated with the vertices i and j, with changes in the predictor associated with i causing changes in the predictor associated with j. In this case, we refer to vertex i as a parent, and to vertex j as a child. When two vertices v and w in a graph are not connected by an edge, the predictors associated with these vertices are conditionally independent. Two predictors I and J are conditionally independent given a third predictor K if given the value of K, the value of I provides no further information about the value of J (see Lauritzen, 1996). Furthermore, graphs in causal inference models tend to be acyclic. In an acyclic graph there is no path from a vertex v through directed edges that leads back to v. A graph with the properties of directionality and acyclicity is referred to as a directed acyclic graphs (DAG).

Finding causal structure in a data set is far from trivial. A first problem is identifiability. The information fed into a causal inference model consists of a measure of association between variables, such as a correlation matrix. As years of cautionary tales in statistical lectures have taught us, however, correlation does not imply causation. The identifiability problem refers to the problem of separating association between variables from causal relationships between variables. A second problem is that the search space increases super-exponentially with the number of predictors (Peters et al., 2014). Even for data sets with a relatively small number of predictors, an exhaustive search of all possible causal structures is therefore often not feasible.

The causal inference model used here is the causal additive model (CAM; Bühlmann, Peters, & Ernest, 2014). The CAM alleviates the computational burden imposed by the size of the search space by decoupling the order search for predictors from the selection of vertices and edges in a DAG. For low-to-medium dimensional data such as the data for the lexical-distributional properties of Mandarin Chinese under investigation here, fitting a causal additive model consists of two steps.

First, the order of the predictors is determined through non-regularized maximum likelihood estimation for an exhaustive set of generalized additive models (GAMs; S. Wood, 2006, 2011) that regress one variable on another (i.e., for the set of possible edges in a causal graph of the data). At each iteration of the algorithm, the edge leading to the largest decrease of the negative log-likelihood is included in the model. Second, the model is pruned through regularization. For each given child, a GAM is fitted with the variables corresponding to each potential parent as a predictor. The edges from all potential parent nodes that do not reach significance in this model at a specific α-level are deemed unnecessary, and are removed from the model.

The result of this fitting procedure is an additive structural equation model that takes the form:

$$X_j = \sum_{k \in pa_D(j)} f_{j,k}(X_k) + \varepsilon_j \qquad (6)$$

with independent $\varepsilon_1, \ldots, \varepsilon_p$, parent nodes $pa_D(j)$ for node j in DAG D, all $\varepsilon_j \sim N(0, \sigma_j^2)$, and smooth functions $f_{j,k}(X_k)$ for predictor $k \in pa_D(j)$.

The value of predictor X_j thus is defined as the sum of smooth functions $f_j, k(X_k)$ for all predictors that are parents of vertex j in the DAG D. An advantage of the use of an additive structural equation model (as opposed to, for instance, a linear structural equation model) is that the underlying causal structure is identifiable from the observational distribution of the variables. Furthermore, CAMs are capable of capturing non-linear causal relations between variables.

We fitted a CAM to the lexical-distributional variables and the naming latencies for the two-character words in Mandarin Chinese using the `CAM` package for R (Peters & Ernest, 2015). We set the α-level for the pruning procedure of the CAM to 0.000001 and set all other parameters to their default values. The output of the CAM algorithm is a p by p adjacency matrix that represents the estimated causal structure in the data. The corresponding DAG provides a visual representation of the fitted CAM.

The naming latencies were averaged across participants prior to analysis. In addition to the variables *Frequency*, *C1 Frequency*, *C2 Frequency*, *C1 Strokes*, *C2 Strokes*, *C1 Entropy*, *C2 Entropy*, and *PMI*, we added two new variables to the data set: *C1 Consistency* and *C2 Consistency*. The variables *C1 Consistency* and *C2 Consistency* are measures of the phonology-to-orthography consistency of the first and second character in a two-character word.

Let p be the pronunciation of character c in the two-character word w and W be the set of words in which p occurs. We define the phonology-to-orthography consistency for word w as the proportion of words in W in which the character that corresponds to p is c. *C1 Consistency* and *C2 Consistency* thus are measures of the extent to which homophony manifests itself with respect to the pronunciation of the first and second character in a two-character word.

The $48,644$ words in the CLD contain $4,895$ unique characters. The total number of unique syllables for the $48,644$ words in the CLD is $1,239$ when tone is taken into consideration. When tone is ignored, this number is reduced to 395. A large number of orthographic units is thus mapped onto a limited set of phonological forms. Consequently, the mapping between phonology and orthography is less than consistent in Mandarin Chinese and homophony is much more widespread in Mandarin Chinese than it is in English.

Previous studies have reported effects of homophony in the lexical decision task (Lee et al., 2015; Wang, Ning, & Zhang, 2012; X. Chen et al., 2009; W. F. Chen, Chao, Chang, & Hsu, 2016), as well as in word naming (Ziegler, Tan, Perry, & Montant, 2000). For the current data set, the effects of *C1 Consistency* and *C2 Consistency* did not reach significance (cf. Y. Liu et al., 2007). We therefore omitted these predictors from the PAMM analysis reported above. As we will demonstrate below, however, the inclusion of these variables in the CAM analysis provides interesting insight into the structural organisation of lexical-distributional space in Mandarin Chinese.

CAM results

The DAG for the lexical-distributional variables and the word naming latencies in Mandarin Chinese is presented in Figure 11. A total of 37 edges reached significance at an α-level of 0.000001. Darker edges indicate stronger causal relations. Edges to the response time (i.e., to the vertex labelled *RT*) are indicates with dashed lines, because the response times in the word naming task are not part of lexical-distributional space.

Consistent with our a priori knowledge about the connection between response times and lexical-distributional variables, response times are connected to the rest of the DAG through incoming edges only: lexical predictors cause changes in naming latencies, but naming latencies do not cause changes in lexical predictors.

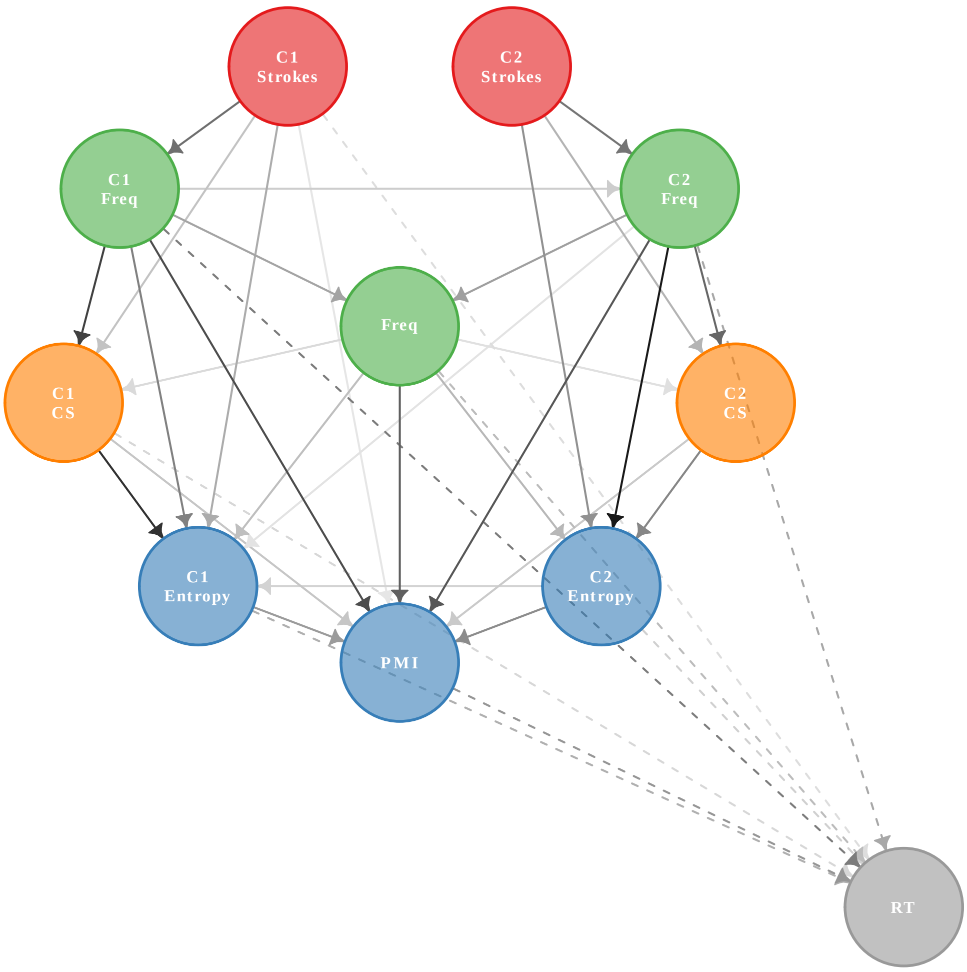

Figure 11. Directed acyclic graph (DAG) of a Causal Additive Model (CAM) fit to the lexical-distributional space in Mandarin Chinese. All edges reached significance at an α-level of 0.000001. Darker edges indicate higher edge scores.

Out of the incoming edges for *RT*, the edge between *C1 Frequency* and *RT* led to the greatest increase in the log-likelihood of the model. Henceforth, we refer to the increase in log-likelihood of the model for an edge as the edge score. The edge score for the edge between *C1 Frequency* and *RT* is 48.156. The strong relationship between *C1 Frequency* and *RT* is in line with the results of the PAMM analysis reported above, in which the effect of *C1 Frequency* had an effect size that was larger than the effect sizes of the effects for the other lexical predictors.

Causal relationships with response times exist for the lexical variables *C2 Frequency* (edge score: 47.583), *Frequency* (edge score: 47.495), *C1 Strokes* (edge score: 47.413), *C1 Entropy* (edge score: 47.542), *C2 Entropy* (edge score: 47.431), and *PMI* (edge score: 47.727) as well. Consistent with the results of the PAMM analysis, no causal relationship between *C2 Strokes* and *RT* was observed at an α-level of 0.000001.

The effects of phonology-to-orthography consistency did not reach significance in the PAMM analysis. Here, we observed a significant causal relationship between *C1 Consistency* (labeled as *C1 CS* in Figure 11) and *RT* (edge score: 47.421), with reduced naming latencies for words with a more consistent phonology-to-orthography mapping for the first character. The causal relationship between *C2 Consistency* (labeled as *C2 CS* in Figure 11) and *RT* did not reach significance.

Below, we discuss the relations between the lexical-distributional variables under investigation. Analogous to our description of the results for *RT* above, we discuss the incoming edges for each lexical-distributional variable. As can be seen at the top of Figure 11, the vertexes for *C1 Strokes* and *C2 Strokes* have no incoming edges. None of the other lexical-distributional variables therefore have a causal impact on these measures of the visual complexity of a character.

C1 Strokes and *C2 Strokes* have a causal influence on the corresponding frequency measures *C1 Frequency* and *C2 Frequency*. The greater the number of strokes in a character, the lower the frequency of that character. The strength of this causal relationship is similar for both characters (edge score *C1*: 48.514; edge score *C2*: 48.297), which suggests that there is a stable causal relationship between the visual complexity and the frequency of a character. This relationship may exist for pragmatic reasons. Characters with fewer strokes may be used more often, because they are easier and less time-consuming to write. Conversely, the increased visual information in characters with a large number of strokes may help readers identify low frequency words more easily (cf Baayen, Milin, Filipović Durdević, Hendrix, & Marelli, 2011; Hendrix, 2016). A causal connection exists between *C1 Frequency* and *C2 Frequency* as well (edge score: 47.438). The higher the frequency of the first character in a two-character compound words, the higher the frequency of the second character. Higher frequency characters thus tend to occur together in compound words. No further incoming edges exist for *C1 Frequency* and *C2 Frequency*.

The frequency of a word is co-determined by the frequency of its component characters, as indicated by significant edges from *C1 Frequency* (edge score: 47.613) and *C2 Frequency* (edge score: −47.645). The higher the frequency of the characters in a word, the higher the frequency of that word. The strength of the causal relationships between the character frequencies and word frequency is moderate. While the frequency of a two-character word depends on the frequency of its component characters to some extent, there are other factors that are not part of the current analysis that influence the frequency of a word as well. The most obvious of these factors is the frequency of the concept a two-character words refers to. The more often the need to refer to this concept arises, the higher the frequency of the word.

As we move down the DAG depicted in Figure 11, the number of incoming edges to each vertex increases. Three incoming edges exist for both *C1 Consistency* and *C2 Consistency*. *C1 Strokes* (edge score: 49.990), *C1 Frequency* (edge score: 47.468), and *Frequency* (edge score: 47.417) have a causal effect on *C1 Consistency*. Similarly, *C2 Strokes* (edge score: 48.774), *C2 Frequency* (edge score: 47.523), and *Frequency* (edge score: 47.410) have a causal effect on *C2 Consistency*. For both characters, higher word and character frequencies lead to a more consistent mapping between phonology and orthography, whereas visually complex characters result in a less consistent mapping between phonology and orthography.

This pattern of results reveals more insight into how the limited set of pronunciations is mapped onto the much larger set of characters. High frequency, visually simple characters have a privileged position in lexical space, in the sense that they own the exclusive rights to dedicated pronunciations. This setup allows language users to access high frequency words as fast as possible. Provided that these words make up the vast majority of word tokens in written language, this reduces overall processing times. By contrast, low frequency and visually complex characters share their pronunciations with other low frequency, visually complex characters. Direct competition for a pronunciation between low frequency words and high frequency words thus is prevented. This helps ensure that access to low frequency characters remains available when it is required.

Significant causal relationships exist between entropy of the first and the second character and all of the corresponding lexical-distributional variables discussed thus far. *C1 Entropy* causally depends on *C1 Strokes* (edge score: 47.561), *C1 Frequency* (edge score: 48.024), *Frequency* (edge score: 47.481), and *C1 Consistency* (edge score: 50.305). Similarly, *C2 Entropy* is influenced in a causal manner by *C2 Strokes* (edge score: 47.481), *C2 Frequency* (edge score: 50.702), *Frequency* (edge score: 47.509), and *C2 Consistency* (edge score: 47.906).

The causal connections between entropy on the one hand and character and word frequency on the other hand are not all that surprising, because both character and word frequency are implicitly (character frequency) or explicitly (word frequency) part of the definition of entropy. High frequency characters occur in a larger number of two-character compound words, which results in a higher entropy. Conversely, if the frequency of a word containing a character is particularly high compared to the frequency of the other two-character that contain the same character, this leads to a reduction in entropy.

The causal relations of stroke counts and phonology-to-orthography consistency with entropy may come as more of a surprise. Characters with more strokes result in less entropy, whereas a more consistent mapping between phonology and orthography causes an increase in entropy. Processing difficulties that arise due to a high degree of visual complexity or an inconsistent phonology to orthography mapping are thus offset by reduced uncertainty about the identity of one character in a two-character word given the other character to ensure identifiability. The picture that emerges then, is that lexical-distributional space in Mandarin Chinese describes a carefully balanced system in which values of one lexical-distributional property that lead to processing difficulties are compensated for by values other lexical-distributional properties that guarantee successful communication.

Furthermore, significant edges from *C2 Frequency* (edge score: 47.408) and *C2 Entropy* (47.426) to *C1 Entropy* exist. The system may thus balance itself both within and across characters. The causal links from *C2 Frequency* and *C2 Entropy* to *C1 Entropy*, however, are relatively weak. Definite conclusions regarding inter-character causal effects, therefore, would be premature.

The intricate interplay between the variables in the DAG that describes lexical-distributional space in Mandarin Chinese culminates at the vertex for *PMI*. This vertex has no less than 8 incoming edges. No causal relation exists between *C2 Strokes* and *PMI* at an α-level of 0.000001. Significant causal connections with *PMI*, however, are present for *C1 Frequency* (edge score: 49.675), *C2 Frequency* (edge score: 49.347), *Frequency* (edge score: 49.059), *C1 Strokes* (edge score: 47.406), *C1 Consistency* (edge score: 47.456), *C2 Consistency* (47.445), *C1 Entropy* (edge score: 47.682), and *C2 Entropy* (edge score: 47.838).

As was the case for the entropy measures, the causal links of *PMI* with *C1 Frequency*, *C2 Frequency*, and *Frequency* are less-than-surprising. PMI depends on all three frequency measures by definition. The frequency of the word is the observed frequency in Equation 3, whereas the expected frequency of a two-character word is a function of the product of the frequency of its component characters. As expected, a greater entropy of the first or the second character furthermore results in a weaker association between both characters. The greater the uncertainty about one character given the other character, the weaker the association between both characters.

The causal relations between PMI on the one hand and the visual complexity and phonology-to-orthography consistency of both characters on the other hand are consistent with the causal connections between these measures and the entropy measures discussed above. As was the case for the entropy measures, visually complex characters and characters with a less consistent phonology-to-orthography mapping lead to higher values of *PMI*. The association between the characters in a two-character words thus is stronger when these characters are more difficult to process individually. The system thus offsets character-level properties that require additional processing with word-level properties that ensure identification of the word as a whole.

The PAMM analysis for the word naming data revealed robust effects of the information-theoretic measures entropy and PMI. During the early stages of the response window, the instantaneous probability of a response is significantly higher for high values of entropy. The effect of PMI persists throughout a larger part of the response window, with a lower probability of a response when the association between the characters in a two-character word is stronger. The CAM analysis provides more insight into both information-theoretic measures and indicates that both measures are embedded in a dynamic system that strives for global optimisation for efficient processing by balancing out local bottlenecks that lead to processing difficulties with lexical-distributional properties that guarantee successful communication. Words with high stroke counts or an inconsistent phonology-to-orthography mapping, for instance, are characterized by reduced uncertainty about the identity of one character given the other character and a stronger association between both characters. Below, we investigate word naming latencies in English to establish if similar principles shape lexical-distributional space and word naming latencies in an alphabetical language.

English

Methods

Participants. For the analyses described below, we use the word naming latencies for compound words in the English Lexicon Project (ELP Balota et al., 2007). Participants for the word naming experiment data in the ELP were recruited from six universities in the United States: Washington University, Wayne State University, Morehead University, University of South Florida, SUNY Albany, and University of Kansas. A total of 443 participants took part in the experiment. The average age of the participants was 23.51 (sd: 9.31).

Materials. The ELP contains word naming latencies for 40,481 words. Each participant named 2,530 words across two experimental sessions that took part within a week of each other. From the set of 40,481 words in the ELP for which word naming latencies are available, we extracted all compound words. This resulted in a data set that contains naming latencies for 2,604 compounds. The data set does not include compounds in which the constituents are separated by a space or a hyphen, as the ELP does not include these types of compounds.

Design. The response variable in our analysis is the average naming latency in the ELP for each compound.

In addition to the control variable *Initial Phoneme*, the predictors in the analyses reported below are lexical-distributional variables that serve as the conceptual equivalent of the lexical-distributional variables in the analysis of the word naming data in Mandarin Chinese reported above: *Frequency*, *Modifier Frequency*, *Head Frequency*, *Modifier Length*, *Head Length*, *Modifier Entropy*, *Head Entropy*, and *PMI*.

We obtained the frequency of the modifier (i.e., the left constituent of a compound), the head (i.e., the right constituent of a compound), and the compound as a whole from the Google 1T n-gram corpus (Brants & Franz, 2006). As was the case for the frequency measures in Mandarin Chinese, we log-transformed each of the frequency measures prior to analysis to remove a rightward skew from the frequency distributions. We thus refer to the frequency measures as *(log) Frequency*, *(log) Modifier Frequency*, and *(log) Head Frequency*.

For the two-character compound words in Mandarin Chinese, stroke counts for the first and second character were included in the analysis as measures of visual complexity. Analogously, we included the length in letters for the modifier and the head as predictors in the analysis of the word naming latencies in the ELP. We applied a square root transform to the length measures to reduce asymmetry in the constituent length distributions. Henceforth, we therefore refer to the length of the modifier and the head as *(sqrt) Modifier Length* and *(sqrt) Head Length*.

We defined *Modifier Entropy* as the entropy over the probabilities of the compounds that share the modifier with the target word (see Table 1). Similarly, *Head Entropy* is defined as the entropy over the probabilities of the compound words that share the head with the target word. We applied a square root transformation to the entropy of the head and the modifier to reduce asymmetry in the entropy distributions. We thus refer to the entropy measures as *(sqrt) Modifier Entropy* and *(sqrt) Head Entropy*.

As was the case for Mandarin Chinese, morphological family size and morphological family frequency are alternative measures of the combinatorial properties of the constituents in a compound in English. Similar to the situation in Mandarin Chinese, the raw correlations of both family size (*(log) Modifier Family Size*: $r = -0.319$, *(log) Head Family Size*: $r = -0.227$) and family frequency (*(log) Modifier Family Frequency*: $r = -0.291$, *(log) Head Family Size*: $r = -0.240$) with the response times are higher than the correlations of the entropy measures with the observed naming latencies (*(sqrt) Modifier Entropy*: $r = -0.257$, *(sqrt) Head Entropy*: $r = -0.194$). To some extent, this is due to the higher correlations of the family size (*(log) Modifier Family Size*: $r = 0.569$, *(log) Head Family Size*: $r = 0.527$) and family frequency measures (*(log) Modifier Family Frequency*: $r = 0.537$, *(log) Head Family Frequency*: $r = 0.523$) with the corresponding constituent frequencies, as compared to the entropy measures (*(sqrt) Entropy Modifier*: $r = 0.441$, *(sqrt) Head Entropy*: $r = 0.469$).

In Mandarin Chinese, the effects of the entropy measures in a linear regression model were more significant than the effects of the measures of family size and family frequency once the frequency of the compound word and its constituents was taken into account. In English, this is not the case. The effects of family size in a linear regression model that includes these frequency counts (*(log) Modifier Family Size*: $t = -10.813$, *(log) Head Family Size*: $t = -11.077$) are somewhat what prominent than the effects of entropy (*(sqrt) Modifier Entropy*: $t = -8.214$, *(sqrt) Head Entropy*: $t = -7.948$), whereas the t-values of the effects of family frequency (*(log) Modifier Family Size*: $t = -7.948$, *(log) Head Family Size*: $t = -8.082$) were similar to those of entropy. Family size thus is a somewhat more powerful predictor for the naming latencies for compound words in the ELP than entropy.

We nonetheless decided to enter the entropy measures rather than the family size measures into the PAMM analysis. The entropy measures and the family size measures are highly correlated. The correlation between *(sqrt) Entropy Modifier* and *(log) Family Size Modifier* is $r = 0.881$, whereas the correlation between *(sqrt) Entropy Head* and *(log) Family Size Head* is $r = 0.899$). To a large extent, the family size and entropy measures thus tap into the same concept. We opted to use entropy rather than family size as a predictor in the PAMM to be able to directly compare the results of the time-to-event analyses in Mandarin Chinese and English. We included both the family size measures and the entropy measures in the CAM analysis of the word naming data for the compounds in the ELP, however.

The third information-theoretic measure under investigation is *PMI*. For English compound words, *PMI* is a measure of the association between the modifier and the head. We define *PMI* as the position-specific point-wise mutual information between the modifier and the head of a compound. Consider, for instance, the compound "stardust". The observed frequency of "stardust" in the Google 1T n-gram corpus is $146,733$. The expected frequency of "stardust" is obtained by applying Equation 2 to the summed frequency of all 6 compounds with the modifier "star" ($1,065,593$), the summed frequency of both compounds with the head "dust" ($401,377$), and the summed frequency of all compounds words ($4,186,689,691$) in the Google 1T n-gram corpus, which

Table 4

Distributional statistics for the predictors (log) Frequency, (log) Modifier Frequency, (log) Head Frequency, (sqrt) Modifier Length, (sqrt) Head Length, (sqrt) Modifier Entropy, (sqrt) Head Entropy, PMI. For each predictor, we provide the original range and the adjusted range after outlier removal, as well as the mean, the median and standard deviation after outlier removal.

predictor	**range**	**adj. range**	**mean**	**median**	**sd**
(log) Frequency	0.00-19.88	0.00-17.43	11.48	11.45	2.03
(log) Modifier Frequency	9.17-22.74	9.17-20.98	16.91	17.02	1.82
(log) Head Frequency	0.00-23.27	0.00-21.56	16.68	16.90	2.39
(sqrt) Modifier Length	1.41-3.00	1.41-2.65	2.03	2.00	0.24
(sqrt) Head Length	1.41-3.46	1.41-2.65	2.08	2.00	0.23
(sqrt) Modifier Entropy	-0.00-1.83	-0.00-1.83	0.97	1.10	0.58
(sqrt) Head Entropy	-0.00-1.89	-0.00-1.89	1.02	1.15	0.63
PMI	-9.39-31.96	-4.43-21.30	8.28	8.28	4.21

yields 102.16. Equation 3 then provides the value of *PMI* for "stardust": $\log_2 \frac{146733}{102.16} = 10.48$. The distribution of *PMI* was near-normal. No transformation was therefore applied prior to analysis.

Consistent with the statistical procedure for Mandarin Chinese, we excluded predictor outliers further than 3 standard deviations from the respective predictor mean from the data prior to the PAMM analysis. As a result, we removed 0.96% of the data (25 words) for *(log) Frequency,*1.96% of the data (51 words) for *(log) Modifier Frequency*, 0.84% of the data (22 words) for *(log) Head Frequency*, 0.46% of the data (12 words) for *(sqrt) Modifier Length*, 1.08% of the data (28 words) for *(sqrt) Head Length*, and 0.69% of the data (18 words) for *PMI*. No outliers further than 3 standard deviations from the predictor mean existed for *(sqrt) Modifier Entropy* and *(sqrt) Entropy Head*. We furthermore removed compound words that started with initial phonemes that occurred less than 10 times in the data prior to analysis. This resulted in the exclusion of 1.19% of the data (31 words). In total, 174 data points (6.68% of the data) were removed prior to analysis. The number of words in the data set for the PAMM analysis thus is 2,430. Table 4 provides distributional statistics for the predictors that entered the PAMM analysis. As before, we present the original ranges of the predictors and the predictor ranges after outlier removal, as well as means, median, and standard deviations of all predictors.

Procedure. Participants in the naming task for the ELP were instructed to respond as fast as possible, while retaining accuracy. Each participant named 2,530 of the 40,481 words across two experimental sessions. Each experimental session consisted of blocks of 250 trials, with a 3 minute pause between blocks. The first session consisted of 1,500 trials, whereas the second session consisted of 1,000 trials. Prior to each trial, three asterisks were shown in the center of the screen and a 50 ms tone was presented to indicate the start of a trial. Next, a word was presented in the center of the screen in the standard QBASIC font in the 80 (column) by 30 (row) mode. The onset of the pronunciation was automatically detected by a voice key. The word remained on the screen for 250 ms after the onset of the pronunciation before the asterisks indicating the start of the next trial appeared on the screen. For more details about the experimental procedure we refer the interested reader to (Balota et al., 2007).

PAMM analysis

The PAMM analysis of the compound word naming data in the ELP was identical to that of the word naming data in Mandarin Chinese. The response time window was limited from 590 ms to 920 ms after stimulus onset. This interval contains 93.70% of the naming latencies in the data. Again, however, it is important to note that the remaining 6.30% of the data remain part of the analysis, despite the fact that the model does not know the exact response times for these words. As before, we divided the response time window into 51 intervals that correspond to the 0, 0.02, 0.04, ..., 0.98, 1.00 quantiles of the naming latency distribution for the PAMM analysis.

The PAMM for the English data models the instantaneous probability of a response as a function of time and the predictors *Initial Phoneme*, *(log) Frequency*, *(log) Modifier Frequency*, *(log) Head Frequency*, *(sqrt) Modifier Length*, *(sqrt) Head Length*, *(sqrt) Modifier Entropy*, *(sqrt) Head Entropy*, and *PMI*. As before, the baseline hazard and time-constant predictor effects were estimated by smooth terms, whereas time-varying predictor effects were estimated with tensor product interactions. Consistent with the PAMM analysis for Mandarin Chinese, we limited predictor smooths as well as time by predictor tensor product terms to fourth order nonlinearities. No restrictions were imposed on the main effect smooth for time that models the baseline hazard.

As was the case for Mandarin Chinese, medium-strength pairwise correlations between the predictors are present. At κ = 67.287, the condition number for the English data set is somewhat higher than the condition number for the Chinese data set (κ = 40.581). As before, we therefore fit separate PAMMs with a main effect smooth for time, a main effect smooth of the predictor and a tensor product interaction between time and the predictor for each of the predictor under investigation. Unless explicitly stated otherwise, the predictor effects in these PAMMs were qualitatively similar to the predictor effects in the full model reported below.

PAMM Results

Table 5 presents the summary of the PAMM fit to the compound word naming data in English. As can be seen in Table 5, the PAMM revealed a significant effect of *time* (χ^2 = 1749.614, p < 0.000) as well as a significant model intercept (z = -52.362, p < 0.000). Together, these terms model the baseline hazard as a function of time, which is presented in Figure 12. The instantaneous probability of a response rapidly increases between 590 and 630 ms after stimulus onset, after which it gradually stabilizes. The functional shape of the baseline hazard is comparable to the functional shape of the baseline hazard in Mandarin Chinese. Consistent with the results for Mandarin Chinese, we observed a significant effect of *Initial Phoneme* as well (χ^2 = 215.512, p < 0.000).

We furthermore found a significant main effect of *(log) Frequency* (χ^2 = 437.554, p < 0.000), as well as a significant interaction of *time* by *(log) Frequency* (χ^2 = 66.235, p < 0.000). The effect of *(log) Frequency* is presented in Figure 13. The instantaneous probability of a response is higher for high frequency compound words from the start of the analysis window (590 ms after stimulus onset) until 850 ms after stimulus onset. The effect of *(log) Frequency* is most prominent at the start of the analysis window. The whole-word frequency effect observed here is similar to the compound frequency effect for Mandarin Chinese reported above. Furthermore, it is in line with the results of previous studies in alphabetical languages that have documented compound frequency effects on response times and eye fixation patterns in a variety of tasks in English (Juhasz, 2008; Andrews, Miller, & Rayner, 2004; De Jong et al., 2002), Dutch (Kuperman, Schreuder, Bertram, & Baayen, 2009; Van Jaarsveld & Rattink, 1988), and

Table 5
Results of the piece-wise exponential additive mixed model (PAMM) fit to the naming latencies. For parametric terms, β estimates, standard errors of the β estimates and p-values are shown. For smooth terms, estimated degrees of freedom, χ^2 values and p-values are provided.

parametric terms	β	**S.E.**	p
Intercept	-5.073	0.097	< 0.001
smooth terms	**edf**	χ^2	p
time	8.865	1749.614	< 0.001
InitialPhoneme	18.137	215.512	< 0.001
(log) Frequency	2.021	437.554	< 0.001
time by *(log) Frequency*	1.002	66.235	< 0.001
(log) Modifier Frequency	2.666	35.341	< 0.001
time by *(log) Mod. Freq.*	2.626	13.781	0.003
(log) Head Frequency	2.901	21.803	< 0.001
time by *(log) Head Freq.*	2.388	8.615	0.076
(sqrt) Modifier Length	2.791	86.155	< 0.001
time by *(sqrt) Mod. Length*	1.483	20.480	< 0.001
(sqrt) Head Length	1.999	29.021	< 0.001
time by *(sqrt) Head Length*	1.873	6.919	0.039
(sqrt) Modifier Entropy	1.000	12.097	0.001
time by *(sqrt) Mod. Entr.*	1.762	3.230	0.259
(sqrt) Head Entropy	1.872	15.010	0.001
time by *(sqrt) Head Entr.*	1.001	0.043	0.836
PMI	1.000	46.120	< 0.001
time by *PMI*	3.840	6.665	0.241

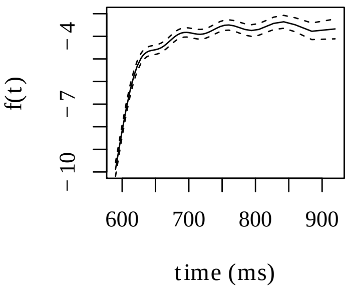

Figure 12. Log-transformed instantaneous hazard function ($f(t)$) with point-wise confidence intervals for English.

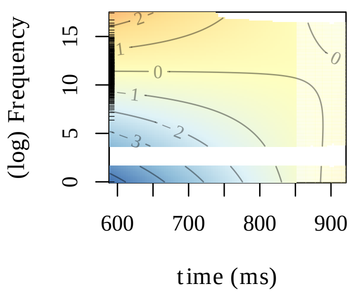

Figure 13. Effect of *(log) Frequency* for English.

Finnish (Kuperman, Bertram, & Baayen, 2008; Pollatsek, Hyönä, & Bertram, 2000).

The frequency of the modifier influences the probability of an instantaneous response as well. The PAMM analysis revealed both a main effect of *(log) Modifier Frequency* ($\chi^2 = 35.341$, $p < 0.000$) and an interaction between *time* and *(log) Modifier Frequency* ($\chi^2 = 13.781$, $p = 0.003$). As can be seen in Figure 14, the effect of *(log) Modifier Frequency* is significant throughout the response time window (590 to 920 ms after stimulus onset), with higher hazard rates for compound words with more frequent modifiers. The effect of modifier frequency is consistent with effects of the frequency of the modifier in eye-tracking studies in English (Andrews et al., 2004; Juhasz, 2008), Dutch (Kuperman et al., 2009; Kuperman, Bertram, & Baayen, 2008), and Finnish (Hyönä & Pollatsek, 1998; Bertram, Hyönä, & Pollatsek, 2004).

Previous studies furthermore reported effects of the frequency of the head of a compound on eye fixation patterns in lexical decision and sentence reading tasks in English (Andrews et al., 2004), Dutch (Kuperman et al., 2009), and Finnish (Kuperman, Bertram, & Baayen, 2008; Pollatsek et al., 2000). The frequency of the head likewise affects the probability of an instantaneous response in the ELP word naming data. We observed a significant main effect of *(log) Head Frequency* ($\chi^2 = 21.803$, $p < 0.000$). The interaction between *time* and *(log) Head Frequency* , however, was marginally significant only ($\chi^2 = 8.615$, $p = 0.076$). The added effect of the significant main effect of *(log) Head Frequency* and the marginally significant interaction of *(log) Head Frequency* with *time* is presented in Figure 15. Consistent with the significant main effect, the effect of *(log) Head Frequency* is significant throughout the analysis window (i.e., from 590 to 920 ms after stimulus onset).

The qualitative nature of effect of *(log) Head Frequency* is more complicated than the qualitative nature

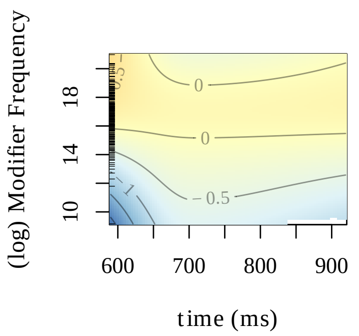

Figure 14. Effect of *(log) Modifier Frequency* for English.

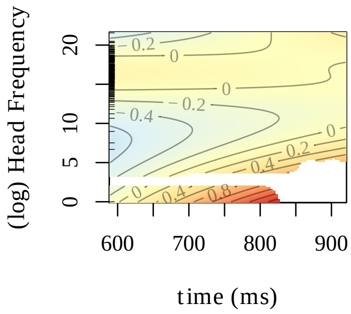

Figure 15. Effect of *(log) Head Frequency* for English.

of the effects we have seen thus far. To some extent, this is due to collinearity. A separate PAMM with a main effect smooth for *time*, a main effect smooth of *(log) Head Frequency* and a tensor product interaction between *time* and *(log) Head Frequency* revealed a somewhat simpler pattern of results. For the early parts of the response window, this model showed a simple facilitatory effect of *(log) Head Frequency*. The early decreased hazard rates for the highest values of *(log) Head Frequency* and the early increased hazard rates for the lowest values of *(log) Head Frequency* therefore do not seem to be robust. The increased hazard rates for compounds with low frequency heads during the later stages of the response window, however, were replicated in the separate PAMM for *(log) Head Frequency*.

Early on, the effect of *(log) Head Frequency* thus is in line with the facilitatory effect of the frequency of the second character reported for Mandarin Chinese above. For compounds that cannot be named during the early stages of the response window, however, the effect of *(log) Head Frequency* reverses. At this point in time, the probability of an instantaneous response is higher for compounds with infrequent heads. For compounds that cannot be responded to during the early stages of the response window uncertainty remains about the identity of the compound during the early stages of processing. A potential explanation for the reversal of the *(log) Head Frequency* effect is that low frequency heads tend to occur in fewer compounds. As noted above, the correlation between *(log) Head Frequency* and *(log) Head Family Size* is $r = -0.227$). Low frequency heads therefore provide more information about the identity of the compound as a whole. This information may help readers reduce the remaining uncertainty during the later stages of the decision making process.

For Mandarin Chinese, we observed an effect of the visual complexity of the first constituent of a two-character compound word only. By contrast, the PAMM for the word naming data in English revealed significant

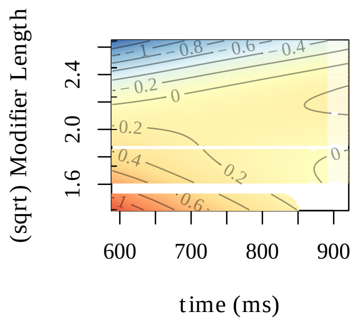

Figure 16. Effect of *(sqrt) Modifier Length* for English.

effects of the visual complexity of both characters. For both *(sqrt) Modifier Length* and *(sqrt) Head Length*, we observed significant main effects (*(sqrt) Modifier Length*: $\chi^2 = 86.155$, $p < 0.000$, *(sqrt) Head Length*: $\chi^2 = 29.021$, $p < 0.000$) that differed significantly as a function of time (*time* by *(sqrt) Modifier Length*: $\chi^2 = 20.480$, $p < 0.000$, *time* by *(sqrt) Head Length*: $\chi^2 = 6.919$, $p = 0.039$).

The effects of the length of the modifier and the head are presented in Figures 16 and 17, respectively. The instantaneous probability of a response is higher for compound words with shorter constituents. The effects of the visual complexity of a compound are most prominent during the early stages of the response window. The effect of *(sqrt) Modifier Length* is less transient than the effect of *(sqrt) Head Length*. The effect of *(sqrt) Head Length* is no longer significant at 744 ms after stimulus onset, whereas the effect of *(sqrt) Modifier Length* remains significant until 893 after stimulus onset. As such, the effect of *(sqrt) Modifier Length* is temporally more widespread than the effect of the visual complexity of the first constituent of a compound word in Mandarin Chinese as well. Whereas the effect of *(sqrt) Modifier Length* persists for 303 ms, the effect of *C1 Strokes* was significant during the first 199 ms of the response window only.

We furthermore observed significant main effects of both *(sqrt) Modifier Entropy* ($\chi^2 = 12.097$, $p < 0.001$) and *(sqrt) Head Entropy* ($\chi^2 = 15.010$, $p < 0.001$). The interaction of both predictors with time, however, did not reached significance *time* by *(sqrt) Modifier Entropy*: $\chi^2 = 3.230$, $p = 0.259$, *time* by *(sqrt) Head Entropy*: $\chi^2 = 0.043$, $p = 0.836$). This suggests that the effects of the entropy of the head and the modifier are relatively stable across time.

The effect of *(sqrt) Modifier Entropy* is presented in Figure 18. Consistent with the effect for Mandarin Chinese reported above, a higher entropy for the left constituent of a compound word corresponds to a higher probability of an instantaneous response. As indicated by the lack of non-faded areas in Figure 18, however, the effect of *(sqrt) Modifier Entropy* fails to reach significance throughout the response time window. This is possible, because nothing prevents the main effect of a predictor and the partial interaction between *time* and that predictor from being opposite in nature. Although this is rare in practice, it is exactly what happens here. The interaction between *time* and *(sqrt) Modifier Entropy* cancels out the significant main effect of *(sqrt) Modifier Entropy* to the extent that it loses significance. The overall effect of *(sqrt) Modifier Entropy* is marginally significant, with a minimum p-value of 0.052 at 637 ms after stimulus onset. The PAMM analysis thus provides some evidence for an effect of *(sqrt) Modifier Entropy*. This evidence, however, is not overwhelming.

We did find robust evidence for an effect of *(sqrt) Head Entropy*. The effect of *(sqrt) Head Entropy* is shown in Figure 19. Consistent with the effect of the entropy of the right constituent in two-character compounds in Mandarin Chinese, the instantaneous probability of a response is higher when the entropy of the head of a compound is high. The effect of *(sqrt) Head Entropy* reaches significance from 628 to 746 ms after

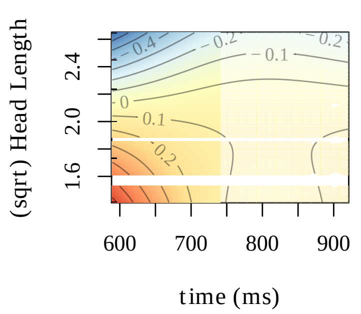

Figure 17. Effect of *(sqrt) Head Length* for English.

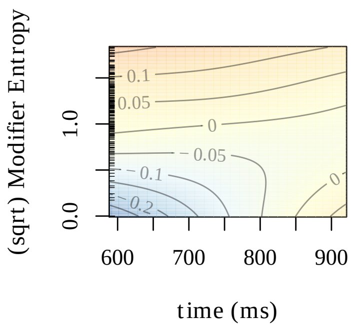

Figure 18. Effect of *(sqrt) Modifier Entropy* for English.

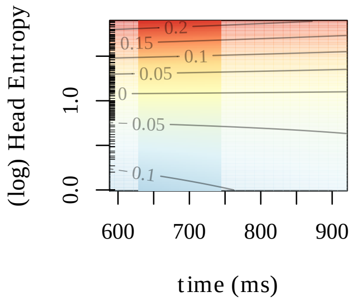

Figure 19. Effect of *(sqrt) Head Entropy* for English.

stimulus onset. As was the case in Mandarin Chinese, the influence of the entropy of the right constituent of a compound on the decision making process thus is greatest during the early stages of the response time window. At these stages, the probability of an instantaneous response is higher for compounds with a head that occurs in more compounds with more similar frequencies.

The association between the constituents in a compound significantly influences the instantaneous probability of a response as well. As was the case for the entropy measures, the main effect of *PMI* was highly significant ($\chi^2 = 46.120$, $p < 0.000$), whereas the interaction between *time* and *PMI* failed to reach significance ($\chi^2 = 6.665$, $p = 0.241$). As was the case in Mandarin Chinese, the effect of *PMI* is less transient than the effects of entropy. As can be seen in Figure 20, the effect of *PMI* is significant from 590 to 821 ms after stimulus onset. The qualitative nature of the effect of *PMI* in English and Mandarin Chinese is similar as well, with a higher instantaneous probability of a response for high values of *PMI* in both languages.

The results for the PAMM analysis in Mandarin Chinese are remarkably similar to the results of the PAMM analysis in English with respect to the effects of the

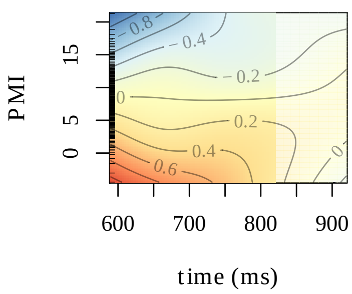

Figure 20. Effect of *PMI* for English.

information-theoretic measures entropy and PMI. The qualitative nature and the temporal profile of the effects of both predictors are similar in both languages. Whereas the effects of the entropy of both characters were highly significant in Mandarin Chinese, however, the effect of entropy of the modifier in English was marginally significant only. This is probably due to the reduced size of the data set for English, as compared to Mandarin Chinese. The effects of entropy are limited to the early stages of the response window. By contrast, the effect of *PMI* remained significant through a large part of the response time window in both languages. Below, we explore the lexical-distributional space for compound words in English in more detail to find out if the similarities between English and Mandarin Chinese exist at the level of relationships between lexical-distributional variables as well.

CAM analysis

We fitted a CAM to the lexical-distributional variables and the naming latencies for compound words in the ELP that was identical to the CAM fit to the data in Mandarin Chinese. As lexical-distributional variables, we included the predictors in the PAMM model reported above: *(log) Frequency*, *(log) Modifier Frequency*, *(log) Head Frequency*, *(sqrt) Modifier Length*, *(sqrt) Head Length*, *(sqrt) Modifier Entropy*, *(sqrt) Head Entropy*, and *PMI*. As noted above, we furthermore added *(log) Modifier Family Size* and *(log) Head Family Size* to the model. As before, we set the α-level for the pruning procedure of the CAM to 0.000001.

CAM results

Figure 21 visualises the DAG for the CAM fit to the word naming latencies and the lexical-distributional variables for compound words in the ELP. At the adopted α-level of 0.000001, 27 edges reach significance. As before, darker edges indicate stronger causal relations, as measured through edge scores (i.e., the decrease in the log-likelihood of the model as a result of including the corresponding edge).

As expected, the vertex for *RT* is connected to the rest of the graph through incoming edges only. A causal relation exists between *RT* and five of the eight lexical-distributional variables: *Modifier Length* (edge score: 129.794), *Head Length* (edge score: 129.579), *Head Frequency* (edge score: 129.381), *Frequency* (edge score: 129.623), and *PMI* (edge score: 130.262). The strong causal relation between *PMI* and *RT* is in line with the strong effect of *PMI* reported in the PAMM analysis above. The edges between *Modifier Frequency*, *Modifier Entropy*, and *Head Entropy* and *RT* did not reach significance at the α-level of 0.000001 adopted here.

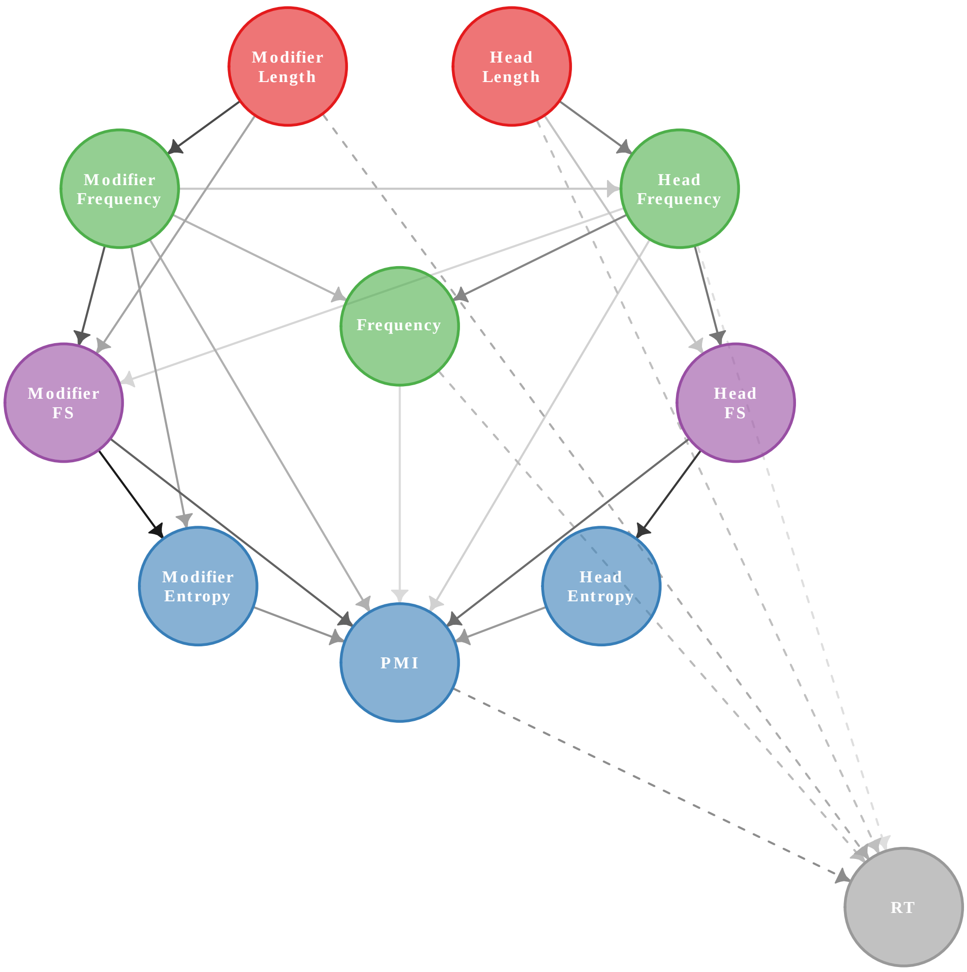

Figure 21. Directed acyclic graph (DAG) of a Causal Additive Model (CAM) fit to the lexical-distributional space in English. All edges reached significance at an α-level of 0.000001. Darker edges indicate higher edge scores.

As was the case for the stroke counts in the CAM for Mandarin Chinese, the vertexes that correspond to the measures of visual complexity *Modifier Length* and *Head Length* in the DAG for English have no incoming edges. The length of the modifier and the head, however, have a causal influence on the respective frequency counts. Consistent with the results for Mandarin Chinese, shorter constituents have higher frequencies. As indicated through the dark edges in Figure 21, the causal connections between *Modifier Length* and *Modifier Frequency* (edge score: 132.769) and between *Head Length* and *Head Frequency* (edge score: 130.637) are strong. Furthermore, as was the case for Mandarin Chinese, the *Modifier Frequency* has a causal influence on *Head Frequency* (edge score: 129.498). The higher the frequency of the modifier, the higher the frequency of the head. In English, too, high frequency constituents thus attract other high frequency constituents. No further incoming edges for *Modifier Frequency* and *Head Frequency* are present in the DAG.

The incoming edges for *Frequency* are identical in English and Chinese as well. Both the frequency of the modifier (*Modifier Frequency*, edge score: 129.674) and the frequency of the head (*Head Frequency*, edge score: 130.442) have a causal influence on the frequency of the compound as a whole. Higher frequency constituents form higher frequency compounds. This is most likely an effect of semantic frequency. High frequency constituents refer to high frequency objects or concepts in the world. These high frequency objects or concepts engage in more frequent relations with other objects or concepts than their low frequency counterparts.

Next, we turn to the first measure of the combinatorial properties of the constituent in a compound: morphological family size. *Modifier Family Size* (labeled as *Modifier FS* in Figure 21) is causally affected by *Modifier Length* (edge score: 129.866), and *Modifier Frequency* (edge score: 132.224). Similarly, causal links exist between *Head Family Size* (labeled as *Head FS* in Figure 21) on the one hand and *Head Length* (edge score: 129.536) and *Head Frequency* (edge score: 130.904) on the other hand. As expected, family sizes are larger for higher frequency constituents. By contrast, longer constituents result in smaller family sizes. To ensure the compounds remain identifiable, the additional processing costs for words with longer constituents are thus balanced out with reduced uncertainty about the identity of one constituent in a compound given the other constituent. Finally, there is a significant causal connection between *Head Frequency* and *Modifier Family Size* (edge score: 129.414). The strength of this connection, however, is relatively weak, as is the raw correlation between *Head Frequency* and *Modifier Family Size* ($r = 0.009$). We therefore do not discuss this edge of the DAG in more detail.

The family size of the modifier and the head have a causal effect on the corresponding entropy measures. The strength of the causal connection between *Modifier Family Size* and *Modifier Entropy* is considerable (edge score: 134.969), as is the causal connection between *Head Family Size* and *Head Entropy* (edge score: 133.397). Unsurprisingly, the entropy is higher for constituents that occur in a larger number of families. As was the in Mandarin Chinese, the CAM furthermore revealed a causal effect *Modifier Frequency* on *Modifier Entropy* (edge score: 129.941), with entropy increasing as a function of modifier frequency.

The number of incoming links to the vertices corresponding to *Modifier Entropy* and *Head Entropy* is lower in English than in Mandarin Chinese. To some extent, this is due to the inclusion of *Modifier Family Size* and *Head Family Size* in the CAM analysis for English. Whereas the visual complexity and the frequency of the constituent have direct causal effects on the entropy of the constituent in Mandarin Chinese, the majority of these causal effects is mediated by the family size measures in the DAG for English. Indeed, a post-hoc analysis revealed the presence of significant causal links from the visual complexity and frequency of the constituents to the entropy measures in a CAM that does not include the family size measures. In addition, the statistical power is the CAM analysis for English is lower than the statistical power of the CAM analysis for Mandarin Chinese, due to the reduced size of the data set. As a result, the relatively weak causal relation between the frequency of the compound and the entropy measures reached significance in Mandarin Chinese, but not in English.

The final lexical-distributional variable, *PMI*, was highly connected to the other lexical variables describing the lexical-distributional space of Mandarin Chinese. The position of *PMI* in the DAG for English is similar. The value of *PMI* depends on 7 of the other 9 lexical variables that were entered into the CAM. Consistent with the lack of a causal relation between *C2 Strokes* and *PMI* in Mandarin Chinese, no significant causal relation with *PMI* is present for *Modifier Length* and *Head Length* at the adopted $alpha$-level of 0.000001. *Modifier Frequency* (edge score: 129.726), *Head Frequency* (edge score: 129.440), *Frequency* (edge score: 129.397), *Modifier Family Size* (edge score: 131.894), *Head Family Size* (edge score: 131.219), *Modifier Entropy* (edge score: 130.145), and *Head Entropy* (edge score: 130.037) all have a causal effect on *PMI*, however. Consistent with the results reported for Mandarin Chinese above, higher values of all of these lexical-distributional variables result in a lower value of *PMI*. The association between the constituents in a compound word thus is higher when the compound and its constituents are more frequent and when the constituents occur in a larger number of compounds with more similar frequencies.

The results of the PAMM and CAM analyses for compound words in Mandarin Chinese and English converge to a remarkable degree. The effect of the visual complexity of the second constituent failed to reach significance in Mandarin Chinese, whereas the effect of the entropy of the first constituent was marginally significant in English. The qualitative nature of the effects of constituent and compound frequency, visual complexity, entropy, and PMI, however, was consistent across both languages. The organisation of lexical-distributional space in both languages showed remarkable similarities as well. The corresponding connection of nearly all connections that reached significance in the CAM for English were significant in the CAM for Mandarin Chinese as well. The DAG for Mandarin Chinese, however, contained a larger number of edges. This is presumably due to the larger size of the data set for Mandarin Chinese, which is a result of the increased prevalence of compound words in this language as compared to English. The image that emerges from the CAM analyses is that the distributional space in both Mandarin Chinese and English describes a complex dynamic system in which properties of compounds words that lead to processing difficulties are offset by lexical-distributional properties that ensure identifiability.

Discussion

We investigated the nature of information-theoretic effects on compound processing in the word naming task in two languages: Mandarin Chinese and English. The word naming data for Mandarin Chinese were obtained through a word naming experiment, in which we asked 3 participants to read aloud 25,935 two-character compound words in Mandarin Chinese. The word naming data for English were extracted from the English Lexicon Project (ELP Balota et al., 2007). For each language, we carried out two analyses of the data: a time-to-event analysis using piece-wise exponential mixed models (PAMMs; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018) and a causal inference analysis using causal additive models (CAMs; Peters et al., 2014).

The aim of the PAMM analysis was to establish to what extent information-theoretic measures influence language processing for compound words in the word naming task. Bien et al. (2005) reported effects of the entropy of the constituents of a compound in a response-association task in Dutch. The entropy effects observed by Bien et al. (2005) were replicated for the word naming data in the current study. For Mandarin Chinese, we observed robust, highly significant effects of the entropy of both the first and the second character on the instantaneous probability of a response. For English, we found evidence for effects of the entropy of both the modifier and the head as well. Compounding, however, is a much less common phenomenon in English than in Mandarin Chinese. The data set for English, therefore, was an order of magnitude smaller than the data set for Mandarin Chinese. As a result, the effect of the entropy of the modifier was marginally significant only. The effect of the head, by contrast, was highly significant.

The qualitative nature of the effects of entropy were similar in Mandarin Chinese and English. The entropy of a constituent is high when that constituent occurs in a larger number of compounds with more similar frequencies. Consistent with the effects of entropy reported for Dutch by Bien et al. (2005), we found facilitatory effects of entropy in both Mandarin Chinese and English. The additional experience with high entropy constituents thus leads to a processing advantage for compounds that contain these constituents. The effects of entropy were most prominent (Mandarin Chinese) or exclusively present (English) during the early stages of the decision making process. The temporal profile of the effect of entropy thus was similar across both languages as well.

We investigated the effects of a second information-theoretic measure as well. Point-wise mutual information (PMI) is a measure of the association between the constituents in a compound word. The more often the constituents occur together and the less often each of the constituents occurs in other compound words, the higher the value of PMI. To our knowledge, effects of PMI have not been reported in previous studies. Here, we observed highly significant effects of PMI on the instantaneous probability of a response in the word naming task in both Mandarin Chinese and English.

As was the case for the effects of entropy, the qualitative nature of the effects of PMI in Mandarin Chinese and English was highly similar. The stronger the association between the constituents in a compound word, the lower the instantaneous probability of a response. The inhibitory effect of PMI might seem surprising. The stronger the association between the constituents in a compound, the more the uncertainty about one constituent given the other constituent is reduced. Conversely, however, readers have more experience with constituents that occur in a larger number of compounds, and that hence are less strongly associated with each of these constituents. Similar to the effects of entropy, the effects of PMI thus may be directly related to readers' experience with the constituents in a compound in the context of derivational words.

The effects of entropy were most prominent during the early stages of the response window. By contrast, the effect of PMI was less transient and remained significant throughout most of the response time window in both Mandarin Chinese and English. Whereas the role of the combinatorial properties of the constituents in a general sense was limited to the earlier stages of the decision making process, the association between the specific constituents in a compound continued to inform this process throughout the response time window. The PAMM analysis thus sheds further light on the temporal profile of the effects of the information-theoretic measures.

English and Dutch are alphabetical languages with a relatively rich inventory of phonological forms, in which complex onsets and offsets exist. Compounding is a relatively infrequency phenomenon in these languages. By contrast, Mandarin Chinese is a logographic language with a limited set of phonological forms, even when the presence of lexical tones is taken into account. To allow language users to efficiently distinguish word forms in the auditory modality, the language heavily relies on compound words. The prevalence of compound words therefore is much higher in Mandarin Chinese than in English. Despite these difference, however, the effects of the information-theoretic measures entropy and PMI on the instantaneous probability of a response in the word naming task are highly similar. The current results, therefore, suggest that the effects of these measures on compound processing may be general property of language processing that is relatively independent of the properties of the language under investigation. Further research, however, is necessary to more solidly establish this conclusion.

The CAM analysis focused on the relations between lexical-distributional variables in both languages. The vertices in the directed acyclic graphs (DAGs) for the CAM models were highly connected in both Mandarin Chinese and English. Lexical-distributional variables thus are not isolated concepts that describe independent properties of compound words and their constituents. Instead, strong causal relations exist between lexical-distributional variables. The values of a lexical predictor thus tend to causally depend on the values of multiple other predictors. The high degree of connectivity in the DAGs reached its apex at the vertices corresponding to the information-theoretic measures entropy and PMI.

The organisation of lexical-distributional space in Mandarin Chinese and English converged to a remarkable degree. As expected, and by definition, the frequency of a constituent had a causal effect on the entropy of that constituent, with higher values of entropy for more frequent constituents in both Mandarin Chinese and English. The entropy of a constituent, however, was also causally affected by its visual complexity, either directly (Mandarin Chinese) or indirectly (English, through the family size of the constituent). Furthermore, less-consistent mappings between the phonology and orthography in Mandarin Chinese results in higher entropies. The variable PMI was likewise causally connected to the other lexical-distributional variables. Both in Mandarin Chinese and in English, the frequency and the entropy of the constituents as well as the frequency of the compound as a whole had causal effects on the association between the characters in a compound word. Furthermore, the consistency of the phonology-to-orthography mapping in Mandarin Chinese had a causal influence not only on the entropy of the constituents, but also on the PMI of the compound as a whole.

The picture that emerges, then, is that the information-theoretic measures entropy and PMI do not constitute isolated concepts that exist independently of other lexical-distributional variables. Instead, strong connections exist between the various lexical-distributional aspects of compound words. Generally speaking, properties of compound words that might lead to processing difficulties are offset by other properties that help correctly identify a compound word. The uncertainty about the identity of one constituent given the other constituent, for instance, tends to be lower for words with a high degree of visual complexity or an inconsistent phonology-to-orthography mapping. Both information-theoretic measures, entropy and PMI thus describe properties of a complex system in which processing impediments are counterbalanced by lexical-distributional characteristics that ensure successful communication. Arguably, Mandarin Chinese is a more engaging showcase of this system as compared to English, due to the higher prevalence of compound words in the languages and the interesting issues that arise due its more limited inventory of phonological forms. Despite the differences between Mandarin Chinese and English, however, the CAM analysis indicates that the way in which the language is optimised for efficient processing and successful communication of compound words is highly similar in both languages.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant number 387774888.

References

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, *16*(1), 285–311.

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.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. I., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity.* New York: Wiley.

Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*, 299–321.

Bender, A., & Scheipl, F. (2018). *pammtools: Piece-wise exponential Additive Mixed Modeling tools.*

Bender, A., Scheipl, F., Hartl, W., Day, A. G., & Küchenhoff, H. (2018). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*. Retrieved from `https://doi.org/10.1093/biostatistics/kxy003`

Bertram, R., Hyönä, J., & Pollatsek, A. (2004). Morphological parsing and the use of segmentation cues in reading Finnish compounds. *Journal of Memory and Language*, *51*, 325-345.

Bien, H., Levelt, W. M. J., & Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences*, *102*, 17876–17881.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram. Version 1.* Philadelphia: Linguistic Data Consortium.

Bühlmann, P., Peters, J., & Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, *42*(6), 2526–2556.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6). Retrieved from e10729.doi:10.1371/journal.pone.0010729

Chen, T. M., & Chen, J. Y. (2006). Morphological encoding in the production of compound words in Mandarin Chinese. *Journal of Memory and Language*, *54*(4), 491–514.

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*.

Chen, X., Hao, M., Geva, E., Zhu, J., & Shu, H. (2009). The role of compound awareness in Chinese children’ s vocabulary acquisition and character reading. *Reading and Writing*, *22*(5), 615.

Chinese Academy of Social Sciences. (2012). *Contemporary Chinese Dictionary* (6th ed.). China: The Commercial Press.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language*, *81*, 555-567.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*, 329-365.

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.

Hendrix, P. (2016). *Experimental explorations of a discrimination learning approach to language processing* (Unpublished doctoral dissertation). Eberhard Karl’s Universität, Tübingen.

Hendrix, P. (2018). A cross-linguistic investigation of response time distributions in lexical decision. *Manuscript.*

Hendrix, P. (2019). A word or two about nonwords: Nonword frequency and semantic neighborhood density effects in the lexical decision task. *Manuscript under revision for Journal of Experimental Psychology: Language, Memory and Cognition.*

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.

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.

Hyönä, J., & Pollatsek, A. (1998). Reading Finnish compound words: Eye fixations are affected by component morphemes. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1612–1627.

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes*, *23*(7–8), 1191–1223.

Juhasz, B. (2008). The processing of compound words in English: Effects of word length on eye movements during reading. *Language and Cognitive Processes*, *23*(7–8), 1057–1088.

Juhasz, B., & Berkowitz, R. (2011). Effects of morphological families on english compound word recognition: A multitask investigation. *Language and Cognitive Processes*, *26*(4), 653–682.

Kuo, W. J., Yeh, T. C., Lee, C. Y., Wu, Y., Chou, C. C., Ho, L. T., … Hsieh, J. C. (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. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*, 1089–1132.

Kuperman, V., Ernestus, M., & Baayen, R. H. (2008). Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. *JASA*, *124*, 3897–3908.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America*, *121*(4), 2261–2271.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental*

Psychology: HPP, *35*, 876–895.

Lauritzen, S. L. (1996). *Graphical models.* Oxford, UK: Clarendon Press.

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., … Hsieh, J. C. (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.

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* (pp. 173–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods*, *39*(2), 192–198.

Ministry of Education of the People's Republic of China. (2013). 通用规范汉字表 *[Table of General Standard Chinese Characters].*

Myers, J. (2006). Processing Chinese compounds: A survey of the literature. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 169–196). Oxford: Oxford University Press.

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* (pp. 65–89). Hillsdale, NJ: Lawrence Erlbaum Associates.

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* (pp. 115–134). Hillsdale, NJ: Lawrence Erlbaum Associates.

Peters, J., & Ernest, J. (2015). CAM: Causal Additive Model (CAM) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=CAM` (R package version 1.0)

Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, *15*(1), 2009–2053.

Peterson, E. (2005). *Mandarin tools: Chinese character dictionary.* (Available through `http://www.mandarintools.com/chardict.html`)

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human, Perception and Performance*, *26*, 820–833.

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: The role of entropy. *Psychonomic Bulletin & Review*, *23*(2), 556–570.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, *19*, 1–30.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sun, C. (2006). *Chinese: A Linguistic Introduction.* United Kingdom: Cambridge University Press.

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (cld): A large-scale lexical database for simplified mandarin chinese. *Behavior Research Methods.* Retrieved from `doi.org/10.3758/s13428-018-1038-3`

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., 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, pp. 59–73). Taipei: Department of Psychology, National Taiwan University.

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.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Van Esch, D. (2012). *Leiden weibo corpus.* (Downloaded from http://lwc.daanvanesch.nl)

Van Jaarsveld, H. J., & Rattink, G. E. (1988). Frequency effects in the processing of lexicalized and novel nominal compounds. *Journal of Psycholinguistic Research*, *17*, 447-473.

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.

Wood, S. (2006). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

Wood, S. (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.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC.

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.

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. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes*, *10*(6), 545–600.

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.