Morphology (2014) 24:407–432
DOI 10.1007/s11525-014-9249-5

Surviving truncation: informativity at the interface of morphology and phonology

Jason A. Shaw · Chong Han · Yuan Ma

Received: 1 June 2014 / Accepted: 18 November 2014 / Published online: 5 December 2014

Abstract When disyllabic words of Chinese are compounded, truncation often applies to yield a disyllabic output that draws one syllable from each of the contributing words, e.g., *xiàdiē* + *fúdù* → ~~*xià*~~*diē* + *fú*~~*dù*~~ → *diēfú*. A substantial portion of new words enter Chinese via this process, and all combinations of four underlying syllables are attested in disyllabic neologisms. Variation in which syllable of the base escapes truncation has been described as arbitrary, but our analysis uncovers a clear pattern. Informative syllables survive. We fit a logistic regression model to Chinese truncation patterns using two indices of informativity, *constituent family size* and *frequency ratio*. Both of these factors were significant predictors. To further explore the nature of informativity-based constraints on surface forms, we analysed the truncation probabilities predicted by the model for each base in the neologism corpus. As truncation probabilities increased so too did model accuracy. We interpret this result as evidence for a constraint regulating the degree of uncertainty in base-truncation mappings.

Keywords Compounds · Chinese · Truncation · Informativity · Corpus studies · Constituent family size · Frequency ratio

Electronic supplementary material The online version of this article (doi:10.1007/s11525-014-9249-5) contains supplementary material, which is available to authorized users.

J.A. Shaw (✉)
MARCS Institute for Brain, Behaviour and Development, School of Humanities and Communication Arts, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia
e-mail: J.Shaw@uws.edu.au

C. Han · Y. Ma
School of Humanities and Communication Arts, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia

1 Introduction

There is a growing body of evidence in support of probabilistic models of word formation whereby gradient patterns in the lexicon have been shown to influence morphological productivity (Albright and Hayes 2003; Hay and Baayen 2005; Plag and Baayen 2009). We apply the data-rich methods of probabilistic linguistics to an analysis of Chinese neologisms, focusing in particular on a truncation pattern that has eluded systematic analysis.

The core data come from a common process of word formation in Modern Standard Chinese whereby the concatenation of two disyllabic words ($2+2=4$) together with truncation ($4 \rightarrow 2$) yields a new disyllabic word. For example, 卫星 *wèixīng* 'satellite', composed of 卫 *wèi* 'defend' + 星 *xīng* 'star' and 电视 *diànshì* 'television', composed of *diàn* 'electric' + *shì* 'look at', are compounded to yield 卫视 *wèishì* 'satellite TV'. The new word takes its first syllable from *wèi*xīng and its second from *diàn*shì. Cases such as this, referred to as metacompounds by Ceccagno and Basciano (2007), result in morphologically complex surface forms that are often semantically opaque unless the underlying disyllabic words are recovered.

Representative examples are provided in (1). The disyllabic words underlying the neologism are shown on the left, followed by the neologism, on the right. In this kind of word formation process, ($2 + 2 = 4 \rightarrow 2$), there are four logically possible truncation patterns that preserve one syllable from each underlying word in the surface form. All of these logically possible truncation patterns are attested and are exemplified in (1).

(1) Four patterns of truncation found in Chinese neologisms

(a)	*kuòdà*	+	*zhāoshōu*	=	*kuòzhāo*
	'expand' 'big'		'beckon' 'bring in'		'expand' 'bring in'
	'enlarge'		'recruit'		'expand admission'
(b)	*shēnqǐng*	+	*jǔbàn*	=	*shēnbàn*
	'state' 'request'		'lift' 'do'		'state' 'do'
	'application'		'conduct (activities)'		'bid for (something)'
(c)	*xiàdiē*	+	*fúdù*	=	*diēfú*
	'down' 'fall'		'width' 'degree'		'fall' 'width'
	'fall'		'range'		'margin of decline'
(d)	*zhōngyāng*	+	yínháng	=	*yānghán*g
	'center' 'core'		'silver' 'row'		'core' 'row'
	'center'		'bank'		'central bank'

In general, patterns of word formation through truncation have been more influential in phonological theory than in morphological theory (Alber and Arndt-Lappe 2012 for overview and discussion). Truncation processes, such as the one exemplified in (1), along with abbreviations, clippings and blends, are sometimes marginalized in morphology in part for their seemingly unsystematic nature. Where theory emphasizes phonological constraints on morphology, as in Prosodic Morphology (McCarthy and Prince 1995b), truncation phenomena have received more attention. Trun-

cation data have contributed, for example, to motivating the Correspondence Theory of Faithfulness (McCarthy and Prince 1999, 1995a; Benua 1995), which has in turn been extended to address a range of morphology-phonology interactions in Optimality Theory (Kenstowicz 1996; Benua 1997; Burzio 2002).

In the particular case of Chinese truncations in (1), the phonological pressure to truncate is straightforward. We presume that truncation in these cases is dictated by a phonological preference for disyllabic words (e.g., Duanmu 1999). The majority of Chinese words are disyllabic, and there are both truncation processes and augmentation processes that conspire to produce them (Duanmu 2012). Our focus is not on the phonological pressures that condition truncation, but rather on the seemingly unsystematic output of the process. The preference for disyllables manifests differently across words. The existence of all logically possible (disyllabic) surface patterns (1a–d) raises the question: what determines the syllables that survive truncation? Some forms preserve the first syllable; some preserve the second. This variation has been described as arbitrary (Ceccagno and Basciano 2007). We pursue the hypothesis that "informativity", the degree to which expanded forms can be predicted from the surface form, dictates which syllables survive truncation.

1.1 Informativity

The informativity hypothesis builds on recent evidence indicating that phonological units with a high degree of information content tend to be emphasized in speech. Informative material tends to be accented (e.g., Calhoun 2010; Pan and Hirschberg 2000; Hirschberg 1993), stressed (Bell and Plag 2013, 2012) and produced with longer duration (Aylett and Turk 2006; Bell et al. 2009). Moreover, informative segments are also less likely to be deleted in connected speech (Priva 2008). Informativity has been used to predict other aspects of phonology systems including the quality of epenthetic vowels (Hume et al. 2013) and the preservation of phonological contrast over time (Wedel et al. 2013a, 2013b). Applied to the data in (1), the basic hypothesis is that the syllables that survive truncation are those that contain the most information about the expanded disyllabic form.

Informativity has been operationalized in a variety of ways. One of the simplest measures is frequency. Frequent words are easier to predict and, therefore, contain less information than infrequent words. There are also semantically-based indices of informativity. Words that have a large number of synonyms have less-specific meanings and therefore less information (Bell and Plag 2012). We adopted two indices of informativity relevant to the specific case at hand. One index, *family constituent size*, quantifies the intuition that, all else equal, under-used material makes a preferable exponent for a new word. We define family constituent size as the number of disyllabic words that share the same right or left member (c.f., Schreuder and Baayen 1997). For example, the forms in (2) share the same left-side member, 绘 ***huì*** **'draw'**. The forms in (3) have the same right-side member, 画 ***huà*** **'drawing'**. Our terminology *left/right member* does not differentiate between the sino-graph, i.e., the Chinese character, or the form-meaning dyad represented by the sino-graph. We are aware of recent work arguing that sino-graphs in Japanese have a deep connection to linguistic

structure (Nagano and Shimada 2014), namely, that they correspond directly to morphomic stems (in the sense of Aronoff 1994). We remain agnostic on this issue for the case of Chinese but adopt, out of convenience, the sino-graph as the unit over which family constituent size is determined. The prediction with respect to informativity is that sino-graphs with large constituent families will be more likely to be truncated in the word formation process described in (1).

The family size prediction can be related to the concept of entropy in information theory (i.e., Shannon 1948).[1] A Chinese lexicon with maximum entropy would have a balanced distribution of sino-graphs across the lexicon such that all sino-graphs have equal family constituent sizes. Under this state of the lexicon, the amount of uncertainty about the underlying word resulting from truncation would be equal for all sino-graphs. There would be no informativity-based reason to select one over the other. The reality of the Chinese lexicon is that family sizes are quite large on average and vary substantially from one sino-graph to the next. Truncating the sino-graph with the larger constituent family (preserving the sino-graph with the smaller family) minimizes uncertainty about the underlying form.

(2) Disyllabic words sharing the same left-side member

绘出	***huì chū***	draw out	**绘**	***huì***	**'draw'**	出	*chū*	'out'
绘制	***huì zhì***	make a drawing	**绘**	***huì***	**'draw'**	制	*zhì*	'make'
绘图	***huì tú***	draw a diagram	**绘**	***huì***	**'draw'**	图	*tú*	'picture'
绘成	***huì chéng***	finish drawing	**绘**	***huì***	**'draw'**	成	*chéng*	'done'
绘画	***huì huà***	draw a painting	**绘**	***huì***	**'draw'**	画	*huà*	'drawing'
绘卷	***huì juàn***	paint a silk scroll	**绘**	***huì***	**'draw'**	卷	*juàn*	'roll'

(3) Disyllabic words sharing the same right-side member

绘**画**	*huì* ***huà***	paint (draw a painting)	绘	*huì*	'draw'	**画**	***huà***	**'drawing'**
绢**画**	*juàn* ***huà***	a silk painting	绢	*juàn*	'silk'	**画**	***huà***	**'drawing'**
图**画**	*tú* ***huà***	a painting	图	*tú*	'picture'	**画**	***huà***	**'drawing'**
裱**画**	*biǎo* ***huà***	frame a painting	裱	*biǎo*	'mount'	**画**	***huà***	**'drawing'**
作**画**	*zuò* ***huà***	paint (make a painting)	作	*zuò*	'draw'	**画**	***huà***	**'drawing'**
版**画**	*bǎn* ***huà***	Imprinting a picture (on clothe, wood, or metal)	版	*bǎn*	'plate'	**画**	***huà***	**'drawing'**

Our other index of informativity is the conditional probability of the whole word given its constituents, or *frequency ratio*. To calculate this, the frequency of a two-sino-graph word is divided by the frequency of a single sino-graph constituent. As with family size, we calculated frequency ratio separately for the left and right members of two-sino-graph words. We exemplify the calculation of frequency ratio with the first disyllabic word in (1c), 下跌 *xiàdiē* 'fall'. This word occurs in the Chinese google books corpus 136,055 times. The first sino-graph of the word, 下 *xià* 'down', occurs in the same corpus 61,229,259 times while the second sino-graph, 跌 *diē 'fall'*, occurs 334,281 times. The frequency ratio for the constituent sino-graphs of 下跌 *xiàdiē* 'fall' are, therefore, 0.0002 (136,055/61,229,259) for 下 *xià* 'down'

[1] For recent applications of entropy to phonology see, e.g., Hall (2012), Hume and Mailhot (2013), Tily and Kuperman (2012); for applications in morphology see, e.g., Milin et al. (2009) and Blevins (2013).

and 0.407 (136,055/334,281) for 跌 *diē 'fall'*. On the informativity hypothesis, the sino-graph 跌 *diē 'fall'* would make a better surface form of 下跌 *xiàdiē* 'fall' by virtue of its higher frequency ratio. This is because 跌 *diē 'fall'* is more predictive of the entire compound than is 下 *xià* 'down'.

1.2 Disyllabic truncation bases in Chinese

Our source of truncation data comes from the *The Contemporary Chinese Dictionary* (Yuan 2002) from which we extracted all 4 → 2 truncations in the neologisms section. The disyllabic words underlying the neologisms are structurally quite diverse. They include compounds of varying types, Verb-Verb compounds, e.g., 侦查 zhēnchá, 'watch-check', Verb-Object compounds, e.g., 结婚 jiéhūn, 'get-marriage', Noun-Noun compounds, e.g, 家庭 jiātíng, 'home-court', and Adjective-Noun compounds, e.g., 高峰 gāofēng, 'tall-peak', as well as "pseudo-compounds". "Pseudo-compound" refers to mono-morphemic Chinese words written with two sino-graphs. Examples include 老虎 *lǎohǔ* 'tiger', 饺子 *jiǎozi* 'dumpling', 筷子 *kuàizi* 'chopsticks'. The last two of these have the same second sino-graph, 子 *zi*, which surfaces in a large number of disyllabic words (it has a left constituent family size of 599) but is often semantically null. Duanmu (2012) describes two types of pseudo-compounds. One type, referred to as XX, is written with two sino-graphs that have similar meaning, e.g., 寒冷 *hanleng* 'cold' 'cold', 弯曲 *wanqu* 'curved' 'round', 图画 *tuhua* 'picture' 'drawing'. The other type, referred to as XO, is written with one sino-graph that represents the meaning of the whole word and another sino-graph that has null meaning, such as 子 *zi* in 饺子 jiǎozi 'dumpling'.

The variation in word types underlying neologisms makes for an interesting test of the informativity hypothesis. In principle, the notion of "informativity" applies to these various word types in the same way. More informative sino-graphs, i.e., sino-graphs with small family sizes and high frequency ratios, are predicted to survive truncation regardless of internal structure. A salient alternative to informativity is that truncation in Chinese abbreviations preserves the non-head constituent (Opper and Sugar 2012). Different compound types in Chinese have different modifier-head structures: AN and NN compounds are right-headed, while the VO compounds are left-headed (e.g., Packard 2000). Coordinative compounds have been analysed as headless (Ceccagno and Basciano 2007). A model based on syntactic constituency would therefore predict that some disyllabic words truncate the first syllable while others truncate the second syllable. From the perspective of the informativity hypothesis, it would be an accident if compounds consistently preserved the non-head member. On the other hand, the informativity-based indices may themselves be surface reflections of grammatical processes, one of which may be non-head preservation. The variation in our corpus allows us to test this hypothesis directly.

To evaluate the role of informativity in conditioning truncation, as well as other factors, we fit a series of logistic regression models to the corpus of truncation patterns extracted from *The Contemporary Chinese Dictionary*. We tested a baseline model, which was blind to the structure of the underlying words and included only

the informativity indices introduced here. We also evaluated how the grammatical factors "headedness" and "compoundhood" contributed to the model. Results indicate that the informativity of the underlying sino-graphs contributes significantly to predicting the surface form of neologisms. We conclude by considering the status of informativity as a macro-parameter minimizing uncertainty between morphological structure and phonological form.

2 Quantifying informativity

The neologism section of *The Contemporary Chinese Dictionary* (Yuan 2002) contains 156 words that were formed through the 4 → 2 truncation process (1). The entire list is provided as supplementary materials. Of the four types of truncation patterns, 46 % were of the type exemplified by (1a); 10 % followed the pattern in (1b); 26 % were of type (1c); and 17 % were type (1d). These percentages are comparable to patterns found in Chinese abbreviations more generally (Opper and Sugar 2012). We focus on the neologisms section of the dictionary to build our investigation around relatively recent products of morphological creativity.

We begin by quantifying indices of informativity for the 312 disyllabic words underlying neologisms. We computed these measures over large corpora of written Chinese including the Google books corpus (Lin et al. 2012; Michel et al. 2011), the Chinese Gigaword corpus (Graff and Chen 2005), and an electronic dictionary of Chinese that we compiled from publically available sources.

In the truncations under study, one sino-graph from each of the underlying words surfaces in the neologism. The informativity prediction is that sino-graphs with smaller constituent families and larger frequency ratios will make better contributions to new words. In reporting the results from our corpus study, we therefore compare the constituent family sizes of sino-graphs that survive truncation with those that do not surface in new words.

2.1 Family constituent size

We calculated sino-graph family constituent size across two different corpora. One was an electronic dictionary of Chinese, which we compiled from two different sources. This was intended to offer a systematic coverage of the Chinese lexicon in computing constituent families. The other was compiled from a corpus of Chinese books. We thought that the book corpus might provide a more valid estimation of the mental lexicon since it may be closer to the type of language sample that a Chinese speaker may encounter. Ultimately, the two corpora converged on very similar family size estimates. We describe both corpora here but for reasons of space report in detail only the family size calculations from the compiled lexicon. Scatter plots of the family size measurements and a brief comparison are provided in the Appendix.

Our compiled lexicon drew words from two electronic sources. One of the sources was a list of the most common Chinese words published by the State Language Commission of China in 2008 (现代汉语常用词表). This list contained 56,064 entries. The second source was a list of mainland vocabulary entitled "duoyuanpinyin

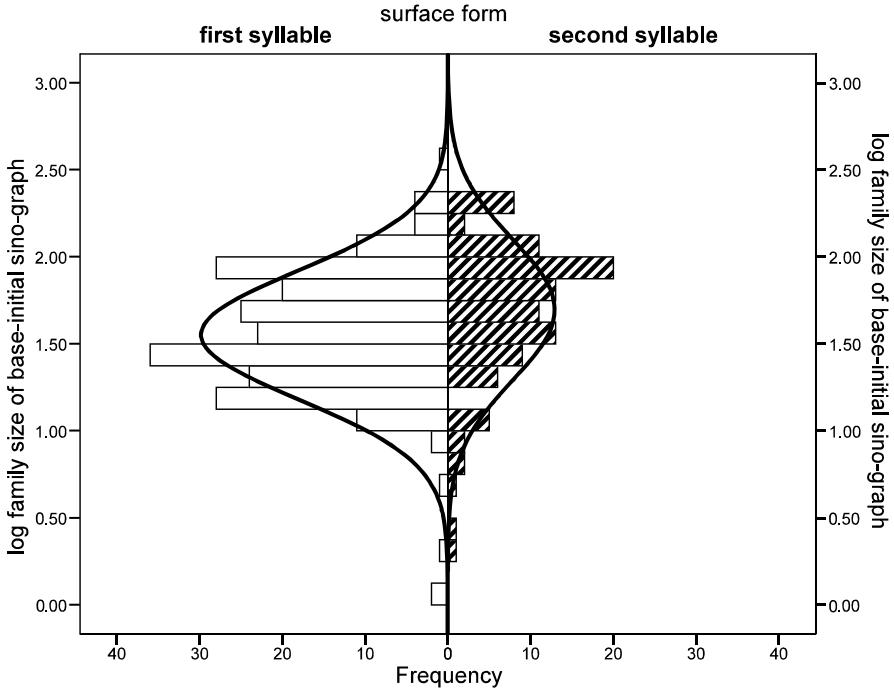

Fig. 1 A pyramid plot of the frequency distribution of *base-initial sino-graph* family sizes for words that project the first syllable (*left*) and words that project the second syllable (*right*)

ciku for richwin" downloaded from http://technology.chtsai.org/wordlist/. This list contained 120,300 entries. The two word lists were combined and duplicate entries were removed. The resulting lexicon contained 132,453 entries. This lexicon was used as one source of calculating the family size of sino-graphs.

In addition to the lexicon-based calculations of family size, we also calculated family size from a large corpus of written Chinese, the Google book corpus accessible via http://books.google.com/ngrams/datasets. The google book corpus is composed of scanned books written in Chinese. The corpus is organized according to *n*-grams. We downloaded all of the 1-grams, i.e., single words, a total of 13,439,617,812 entries.

2.1.1 The left sino-graph of underlying compounds

Figure 1 shows a pyramid plot of family constituent size for the first sino-graph of each underlying word in our neologism corpus ($N = 312$). This plot includes the family size for words such as *kuòdà* 扩大, *shēnqǐng* 申请, *xiàdiē* 下跌, and *zhōngyāng* 中央, shown in (1), which contribute the first sino-graph to the neologisms *kuòzhāo* 扩招, *shēnbàn* 申办, *diēfú* 跌幅, and *yāngháng* 央行, respectively. The family constituent sizes reported here were those calculated over our compiled Chinese lexicon (132,453 entries). The *y*-axis shows the log family constituent size of the first

sino-graph. The centreline divides the family size calculations based on two groups of disyllabic words, those that project the first syllable (left side), e.g., *kuòdà*, *shēnqǐng*, and those that project the second syllable (right side), e.g., *xiàdiē*, *zhōngyāng*, into neologisms. The *x*-axis shows the frequency of each family size bin in the corpus. For reference, the solid black line shows a normal distribution fit to the frequency bins. If truncation was blind to the family size of the first sino-graph, then the distribution of log family sizes on the left and the right of the pyramid plot should not differ.

It is clear from Fig. 1 that the family size distributions are different for words that project the first syllable than for words that project the second syllable. The mean family size, indicated by the peaks of the normal distribution curves, is smaller on the left side of the figure than on the right. This indicates that words that project the first syllable (and truncate the second syllable) tend to have smaller constituent families for the sino-graph representing the first syllable than words that project the second syllable (and truncate the first syllable). These results are in the predicted direction of the informativity hypothesis. Sino-graphs with smaller family sizes carry more information about the underlying compound than sino-graphs with larger families. It is therefore easier to recover a disyllabic base from a sino-graph with a small family than from a sino-graph with a large family.

2.1.2 The right sino-graph of underlying compounds

Figure 2 shows the frequency distribution of family size for the second, or base-final, sino-graph of disyllabic words. This includes, for example, *zhāoshōu* 招收, *jǔbàn* 举办, *fúdú* 复读, and *yínháng* 银行, words that contribute, respectively, the second syllable to neologisms: *kuòzhāo* 扩招, *shēnbàn* 申办, *diēfú* 跌幅, and *yāngháng* 央行 (see (1)). The format of the figure follows Fig. 1. The family size of disyllabic words that project the first syllable into neologisms (truncating the second syllable) are shown on the left and the family size of words that project the second syllable (truncating the first syllable) are shown on the right. The *y*-axis shows log family size. The *x*-axis shows the frequency of each family size bin. The solid black line shows the normal distribution.

Figure 2 also shows constituent family size distributions that are consistent with the informativity hypothesis. The mean family size of the second sino-graph is higher for words that project first sino-graph (left side of Fig. 2) than for words that project the second sino-graph (right side of Fig. 2).

Taken together, Figs. 1 and 2 demonstrate a clear trend. Sino-graph family size is related systematically to syllable truncation in new word formation. Specifically, sino-graph family size tends to be larger for syllables that get truncated.

2.1.3 The relation between dictionary and corpus-based calculations of family size

In addition to family size calculations based on our compiled lexicon, we also calculated constituent families based upon a large corpus of written Chinese. The results from the written corpus produced highly similar family-sizes to those summa-

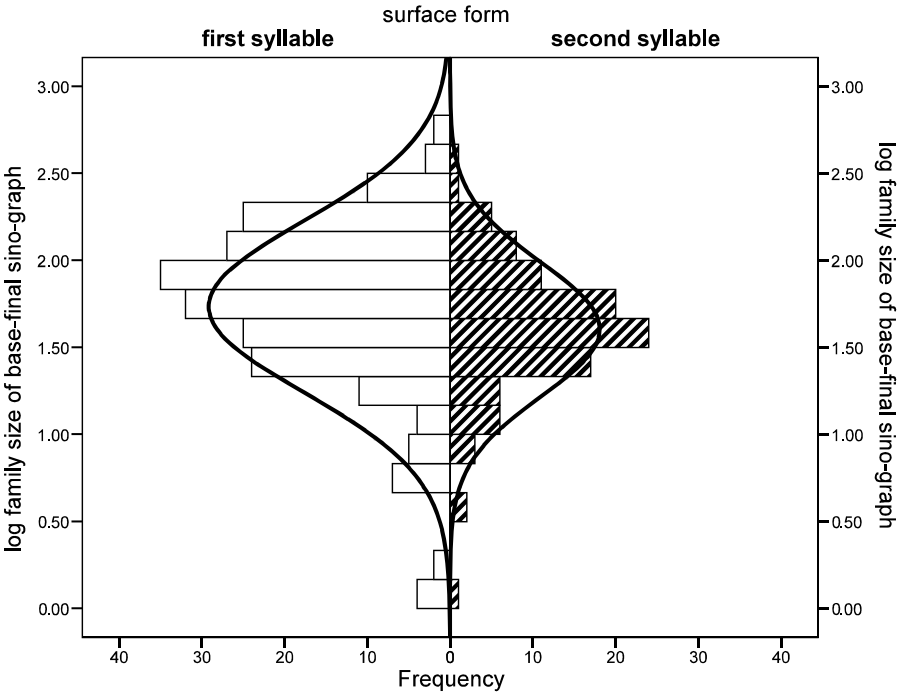

Fig. 2 A pyramid plot of the frequency distribution of *base-final sino-graph* family sizes for words that project the first syllable (*left*) and words that project the second syllable (*right*)

rized above for the lexicon-based calculations. Overall, the two calculations of family size are highly correlated for both the first sino-graph of compounds, $r = 0.803$; $p < 0.001$, and the second, $r = 0.881$; $p < 0.001$. The Appendix shows scatter plots of the two measures and includes a brief discussion of the differences.

2.2 Frequency ratio

The second index of informativity employed in this study was the conditional probability of the compound given the constituent sino-graph or *frequency ratio*, defined as the frequency of the whole disyllabic word relative to the frequency of each individual sino-graph. Frequency ratio was calculated separately for both the left and right sino-graphs of underlying disyllabic words. The frequency of the component sino-graphs and the frequency of disyllabic words were calculated in two independent corpora. The first corpus was the subset of the Chinese Gigaword corpus (LDC2005T14) drawn from the Xinhua News Agency (Graff and Chen 2005). This corpus contains 992,261 sino-graphs segmented into 311,660 words. The second corpus over which frequency ratios were calculated was the Google Chinese books corpus, described above. In reporting frequency ratios, we focus on the Chinese Gigaword corpus, but we offer a comparison of the two corpora in the Appendix.

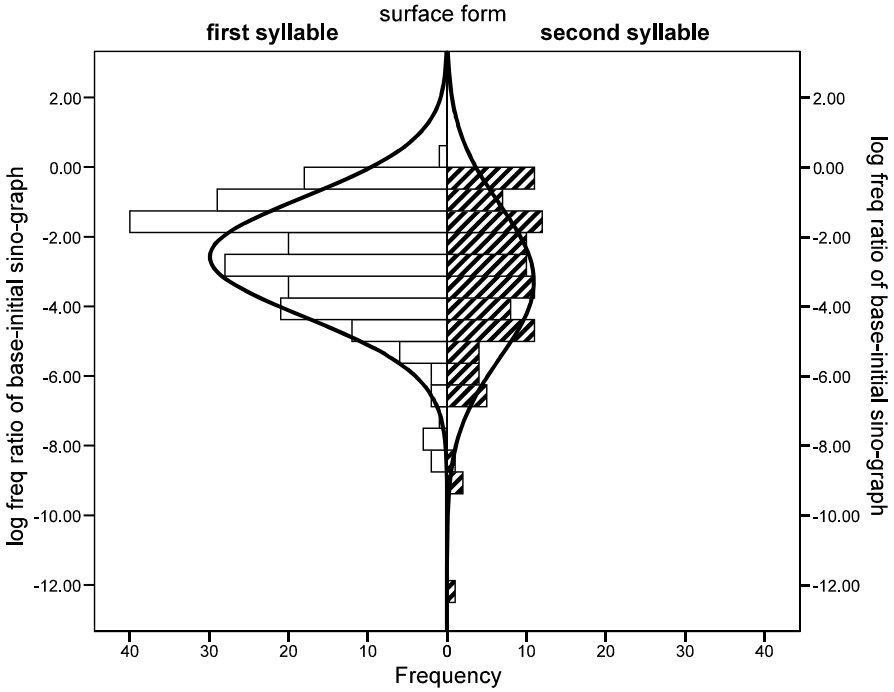

Fig. 3 Log frequency ratio of the *base-initial sino-graph* for words that project the first syllable (*left*) and words that project the second syllable (*right*)

2.2.1 Frequency ratio of the first sino-graph

Figure 3 shows a pyramid plot of frequency ratio for the first sino-graph of each underlying word in our neologism corpus ($N = 312$). Referring back to (1), these are, for example, *kuò* in *kuòdà*, *shēn* in *shēnqǐng*, *xià* in *xiàdiē*, and *zhōng* in *zhōngyāng*, which project a sino-graph into the neologisms *kuòzhāo*, *shēnbàn*, *diēfú*, and *yāngháng*, respectively. The frequency ratios were calculated from the Xinhua news sub-section of the Gigaword corpus. The *y*-axis shows the log frequency ratio of the first sino-graph. As with the pyramid plots of family constituent size reported above, the centreline divides the frequency ratio measurements into groups based upon the sino-graph that survives truncation in new words. Underlying compounds that project the first syllable are on the left side of the figure. Sino-graphs that project the second syllable are shown on the right side of the figure. The *x*-axis shows counts of each log frequency ratio bin. The solid black line shows a normal distribution fit to the frequency bins. If truncation was blind to the frequency ratio of the first sino-graph, then the distribution of log frequency ratios on the left and the right of the pyramid plot should not differ.

It is clear from Fig. 3 that the distribution of frequency ratios are different for words that project the first syllable than for words that project the second syllable. The frequency ratios tend to be larger on the left side of the figure than on the right. This indicates that words that project the first syllable (and truncate the second sylla-

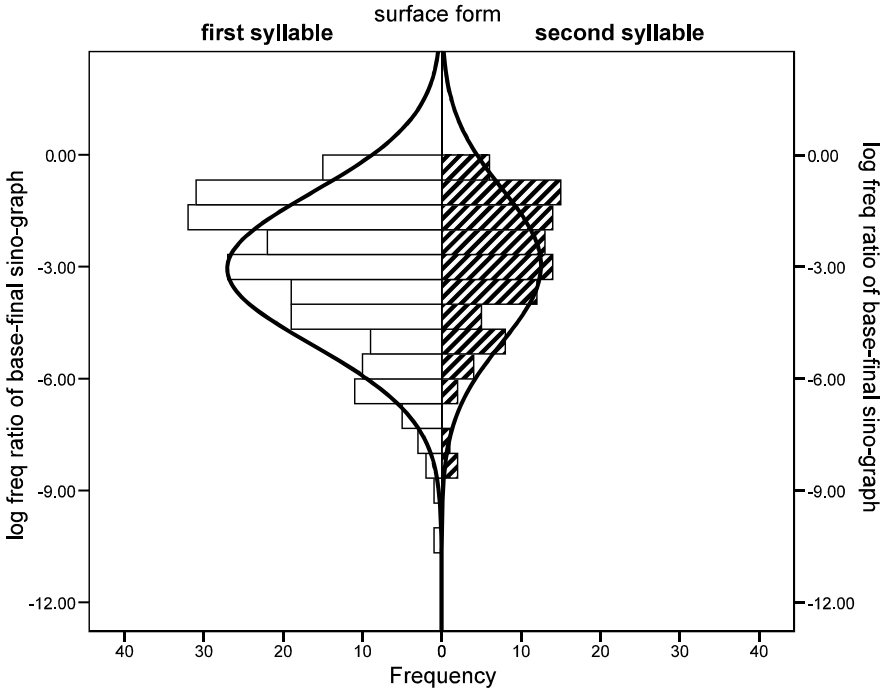

Fig. 4 Log frequency ratio of the *base-final sino-graph* for words that project the first syllable (*left*) and words that project the second syllable (*right*)

ble) tend to have larger frequency ratios than words that project the second syllable (and truncate the first syllable). These results are also in the predicted direction of the informativity hypothesis. Sino-graphs with high frequency ratios carry more information about the underlying di-syllable than sino-graphs with smaller frequency ratios. It is therefore easier to recover the disyllabic word from a sino-graph with a high frequency ratio than from a sino-graph with a low frequency ratio.

2.2.2 Frequency ratio of the second sino-graph

The informativity hypothesis predicts that the frequency ratio of the second sino-graph mirrors that of the first sino-graph. For disyllabic words that project the first sino-graph, second syllable frequency ratio should be low. For disyllabic words that project the second sino-graph, second syllable frequency ratios should be high.

Figure 4 shows a pyramid plot of frequency ratio for the second (base-final) sino-graph of underlying disyllabic words. The format of the figure is the same as Fig. 3 and the frequency ratios were calculated in the same corpus, the Xinhua News subsection of the Gigaword corpus. In line with the informativity hypothesis, frequency ratios are higher on the right side of the figure than on the left side of the figure. However, the difference is very small, particularly when compared to pyramid plots of family size for first (Fig. 1) and second (Fig. 2) sino-graphs and to the pyramid plot of the frequency ratio of the first sino-graph of underlying di-syllables (Fig. 3).

2.2.3 The relation between frequency ratio calculations across corpora

In addition to the Gigaword corpus, we also calculated frequency ratio using the Google books corpus. The results were highly similar. Frequency ratios from the two corpora were highly correlated for both first sino-graph ($r = 0.820$; $p < 0.001$) and second sino-graph ($r = 0.881$; $p < 0.001$) frequency ratios. Scatter-plots of frequency ratios across corpora and a brief discussion are provided in the Appendix.

In summary, frequency ratio provides another index of informativity that potentially influences new word formation. The sino-graphs that survive truncation in our corpus tend to be those that have higher frequency ratios. Sino-graphs with high frequency ratios contain more information about the base than sino-graphs with low frequency ratios. Both frequency ratio and family size showed stability across corpora.

3 Statistical models of truncation

Comparing indices of informativity across truncated and preserved sino-graphs reveals two main trends. Sino-graphs that survive truncation tend to have small morphological families and high frequency ratios. Such sino-graphs provide the maximum amount of information about the underlyingly disyllabic form of the word being truncated. To evaluate the statistical significance of these trends, we fit a binary logistic regression model to the data.

3.1 The informativity-based statistical model

All disyllabic words underlying $4 \rightarrow 2$ truncations were coded for whether the first syllable (the default, coded as 0) or the second syllable (the less common pattern, coded as 1) surfaced in neologisms. The informativity indices discussed above were used to predict the log odds of the second sino-graph surfacing. The model was implemented using the *lrm* function from package *rms* (version 4.0) in *R* (version 3.0.2). Frequency ratios were log-transformed to better approximate a normal distribution.

Since the informativity indices calculated over different corpora were highly correlated, $r > 0.8$ for all indices (see also Appendix), they cannot be included as independent predictors. Therefore, separate models were fit to each set of measurements. Table 1 offers a comparison of the different models, including the concordance index, or C-index, and the χ^2 statistic. The C-index provides an index of how well model predictions match the data—a value of 0.5 indicates random predictions; a value of 1 indicates perfect predictions (Baayen 2008:181).

We tried all combinations of frequency ratio and family size calculations from the different corpora. The model that accounted for the most variance in the data and provided the most accurate predictions used the dictionary to compute family sizes and the Gigaword corpus to compute frequency ratio. This result was somewhat surprising, since the Google books corpus is an order of magnitude larger than the Gigaword corpus. This difference may be attributable to the quality of segmentation in the respective corpora. The Gigaword corpus is older and has therefore been subjected to

Table 1 The concordance index and, in parenthesis, the χ^2 statistic for models based on different calculations of family size and frequency ratio. '***' indicates significance at $p < 0.001$. The combination that accounted for the most variance in truncation patterns involved family size computed over the dictionary and frequency ratio calculated across the Chinese Gigaword corpus

Model performance: C-index (χ^2)		Frequency ratio	
		Gigaword	Google
Family size	Dictionary	**0.693(34.13***)**	0.691(28.71***)
	Google	0.679(28.84***)	0.679(25.43***)

Table 2 Summary of logistic regression model of truncation patterns based on informativity

	β	S.E.	Wald Z	Pr (>|Z|)
Intercept	−1.1575	0.3073	−3.77	0.0002
Family size, Sino-graph 1	0.0068	0.0027	2.56	**0.0104***
Family size, Sino-graph 2	−0.0065	0.0026	−2.46	**0.0137***
Frequency ratio, Sino-graph 1	−0.2574	0.2457	−2.41	**0.0158***
Frequency ratio, Sino-graph 2	0.0956	0.2497	0.88	0.3779

greater academic scrutiny. Although there have been recent advances in segmenting Chinese (see, e.g., Sproat and Shih 2002 for a tutorial), there may be a trade-off between the size of the corpus and the quality of the segmentation. The Appendix provides a comparison of informativity measurements across corpora. In the remainder of the models, we use constituent family sizes calculated from the dictionary and frequency ratios calculated from Gigaword.

We assessed the collinearity of the four indices of informativity (first sino-graph family size, second sino-graph family size, first sino-graph frequency ratio, second sino-graph frequency ratio) for the best fitting model. These predictors are largely orthogonal. There were no pairwise correlations greater than $r = 0.02$. The condition number, κ, used to assess collinearity was 4.64. Values of κ between 0 and 6 indicate that there is no collinearity to speak of (Baayen 2008:200).

The best fitting informativity-based model is summarized in Table 2. The direction of the effects, determined by the sign of the β coefficient, supports the informativity hypothesis for all predictors. The positive coefficient for the family size of the first (leftmost) sino-graph indicates, in line with predictions, that the odds of the second syllable surfacing are positively correlated with the family size of the first sino-graph. As first sino-graph family size increases, the likelihood of having the second sino-graph surface goes up. The negative sign of the coefficient for second (base-final) sino-graph family size is also in line with predictions. The family size of the second sino-graph varies inversely with the odds of the second sino-graph surfacing. The second sino-graph is less likely to surface in neologisms when it has a large family size. The signs of the coefficients for frequency ratio are also in line with informativity predictions. The negative sign for first sino-graph frequency ratio indicates that the odds of the second sino-graph surfacing go down as the frequency ratio of the first sino-graph goes up. Lastly, the coefficient for second sino-graph frequency ratio

is positive, as predicted. This indicates that the second sino-graph tends to surface more often when it has a high frequency ratio. Thus, the direction of each effect is as predicted by the informativity hypothesis. Of the four predictor variables, however, only three accounted for a significant amount variance. Family size was significant for both the first and the second sino-graph of the base, but frequency ratio was only significant for the first sino-graph.

The model accounts for a significant portion of variance in the truncation data ($\chi^2 = 36.89$, $p < 0.0001$). The effects of the informativity predictors echo trends displayed in pyramid plots (Sect. 2) providing support for the hypothesis that informativity influences truncation patterns in Chinese compounds.

3.2 The role of headedness and compoundhood in truncation

We next evaluate whether headedness accounts for variance in the data above and beyond informativity. Since the notion of headedness is not relevant for pseudo compounds, we first subset the larger corpus into pseudo-compounds and real compounds. Of the 312 words underlying neologisms, 176 were pseudo-compounds leaving us with 136 real compounds. The sub-corpus of real compounds was coded for three levels of headedness: Adjective-Noun compounds (and some NN compounds) were coded as right-headed; Verb-Object compounds were coded as left-headed; Verb-Verb compounds and coordinative Noun-Noun compounds were coded as headless (Ceccagno and Basciano 2007).

We fit logistic regression models to the compound subset (leaving out the pseudo compounds). The first model contained only headedness (right, left, none) as a factor. By itself, headedness was not a significant predictor of truncation in the compounds sub-corpus (C-index = 0.51; $\chi^2 = 0.02$, $p = 0.87$). When added to the baseline informativity model, containing family size and frequency ratio, headedness provides a slight increase in variance explained, e.g., $\chi^2 = 14.60$ without headedness; $\chi^2 = 15.41$ with headedness in the model, but does not improve on the C-index: without headedness = 0.696; with headedness = 0.695. The negligible increase in variance explained does not warrant increased model complexity. The Akaike Information Criterion (AIC), which summarizes goodness of fit while penalizing model complexity (Bozdogan 1987), is lower without the headedness parameter included in the model (AIC without headedness = 158.04 vs. AIC with headedness = 160.03), indicating that the best model of truncation in real compounds is one that excludes compound headedness.

The model of real compounds ($n = 136$) excluding headedness is summarized in Table 3. All effects are in the direction predicted by the informativity hypothesis, as in the model of the entire 312 word corpus (Table 2). However, the size of the effects is different. The influence of family size has diminished substantially and is no longer a significant predictor while the effect of frequency ratio is enhanced. The lower p values are not particularly surprising since the real compounds subset is less than half the size of the entire corpus. It is also possible, however, that the effect of our informativity indices on truncation interacts with word type. To assess this possibility, we also included compoundhood (pseudo- vs. real compounds) and interactions between compoundhood and informativity indices as predictors in a model of the entire corpus ($n = 312$). Real compounds were coded as 1; pseudocompounds were coded as 2.

Table 3 Summary of logistic regression model of truncation patterns for real compounds only ($n = 136$)

Model of truncation in real compounds				
Real compounds	β	S.E.	Wald Z	Pr (>|Z|)
Intercept	−0.8031	0.4864	−1.65	0.0987
Family size, Sino-graph 1	0.0022	0.0035	0.62	0.5342
Family size, Sino-graph 2	−0.0046	0.0035	−1.34	0.1787
Frequency ratio, Sino-graph 1	−0.4110	0.1623	−2.53	**0.0113**
Frequency ratio, Sino-graph 2	0.2374	0.1522	1.56	0.1188

Table 4 Summary of logistic regression model of truncation patterns for real compounds only ($n = 312$)

Pseudo compounds	β	S.E.	Wald Z	Pr (>|Z|)
Intercept	−0.3198	0.5768	−0.55	0.5793
Compoundhood (pseudo vs. real)	−0.4546	0.2746	−1.66	0.0979
Family size, Sino-graph 1	0.0064	0.0027	2.37	**0.0178**
Family size, Sino-graph 2	−0.0079	0.0029	−2.73	**0.0063**
Frequency ratio, Sino-graph 1	−0.2603	0.1076	−2.42	**0.0156**
Frequency ratio, Sino-graph 2	0.1082	0.1098	0.99	0.3243

The model including compoundhood is summarized in Table 4. Although the effect was not significant, including compoundhood improved the model beyond the baseline informativity model in Table 2: with compoundhood, C = 0.702; $\chi^2 = 36.89$; AIC = 368.00 vs. without compoundhood C = 0.693; $\chi^2 = 34.13$; AIC = 368.76. The direction of the effect (negative β coefficient) indicates that pseudo compounds are less likely to preserve the second sino-graph than real compounds. The direction and magnitude of the informativity-based effects are not altered by the addition of compoundhood to the model. Adding interactions between compoundhood and informativity indices led to further improvements in C = 0.71 and $\chi^2 = 42.68$ but AIC increased to 370.21, indicating that the additional interaction terms are not explaining enough variance to justify inclusion. Thus, the tendency for pseudo-compounds to preserve the first sino-graph is largely independent of informativity.

To sum up the statistical analysis, all four informativity-based factors (family size of first sino-graph, family size of second sino-graph, frequency ratio of first sino-graph, frequency ratio of second sino-graph) influenced truncation in the predicted direction. Across the whole corpus ($n = 312$), three of these factors accounted for a significant amount of variance. Only the frequency ratio of the second sino-graph did not reach significance. We also added "headedness" to the model and tested it on the real compounds ($n = 136$). This factor did not account for truncation patterns by itself or in conjunction with other factors. Lastly, as the majority of disyllabic words were pseudocompounds ($n = 176$), we tested the effect of compoundhood (pseudo vs. real). This factor improved the model, as there is a mild tendency for pseudo-compounds to preserve the first sino-graph. The interaction between compoundhood

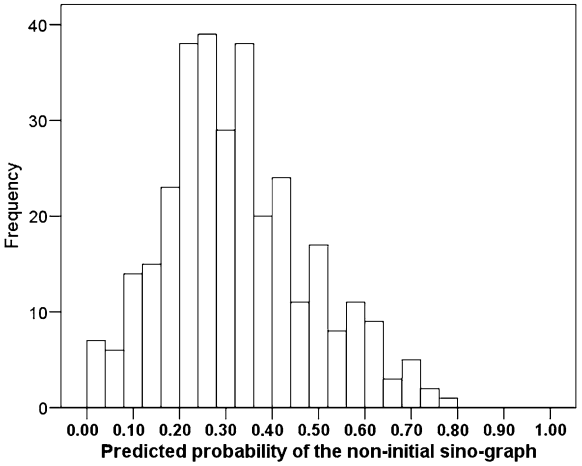

Fig. 5 The predicted probability of the second sino-graph surfacing in neologisms

and other informativity indices was not significant and did not improve the model, indicating that the effect of compoundhood is independent of informativity.

3.3 Predicted probabilities

To explore informativity predictions in greater detail, we used the regression coefficients from our baseline model (Table 2) in a generative fashion, predicting the probability of sino-graph two, the non-initial sino-graph of the base, surviving truncation for each underlying compound in our corpus. Following the equation in (2), the probability of sino-graph two, Char2, surfacing in a neologism was defined as a function of the informativity indices: family size of sino-graph one (fs_1), family size of sino-graph two (fs_2), the frequency ratio of sino-graph one (fr_1) and the frequency ratio of sino-graph two (fr_2). The exponential term in the equation is a linear function of the informativity measures weighted by the regression coefficients. The equation was applied to the informativity indices of the neologisms corpus to predict truncation probabilities for each word.

$$(4) \qquad P(char2|fs_1, fs_2, fr_1, fr_2) = \frac{1}{1 + e^{\beta_0 + \beta_1 fs_1 + \beta_2 fs_2 + \beta_3 fr_1 + \beta_4 fr_2}}$$

Figure 5 shows a histogram of the predicted probabilities. The bulk of the predictions are to the left of the 0.50 mark on the x-axis. This indicates that the model tends to predict that the first sino-graph survives truncation. This general pattern, first sino-graph surviving truncation more often than the second, is upheld in the data, and has been described more generally as the "default" truncation pattern for Chinese abbreviations with the alternative pattern, second sino-graph surfacing considered "exceptional" (Opper and Sugar 2012). As Fig. 5 shows, the "default" pattern is a prediction of the informativity hypothesis for the Chinese compounds considered here.

The values of the tails of the distribution are also informative. The right tail of the distribution stops at 0.80. This indicates that the second sino-graph is never predicted to surface at greater than 0.80 probability. In contrast, the left-edge of the distribution stretches all the way to zero. This indicates that there is something of an island of

reliability, in the sense of Albright (2002), within the "default" pattern. The first sino-graph is predicted to surface with probabilities of greater than 0.80 (values of zero to 0.20 in Fig. 5) for a substantial number of compounds.

The predictions in Fig. 5 can be evaluated on a case-by-case basis. This requires interpreting the probabilities output by the model in terms of discrete predictions. A typical way to do this in the regression literature is to use 0.50 as a threshold for prediction. Probabilities less than 0.50 predict the first sino-graph; probabilities greater than 0.50 predict the second sino-graph. The 0.50 threshold tests model performance as a categorical predictor in a forced choice type of task. Harrell (2001:248–249) enumerates several reasons why, as a general rule, the 0.50 threshold can be misleading as a metric of model performance. He advocates instead the use of the C-index, which we reported for our baseline model in Table 1. The C-index, as applied to our informativity-based model, considers all word pairs in which one word projects the first sino-graph (truncating the second) and the other projects the second sino-graph (truncating the first). The index is the proportion of such pairs in which the predictions are in the correct direction, i.e., cases in which the sino-graph with a higher predicted probability actually surfaces. For our model, as it turns out, the C-index provides a similar result as the 0.50 threshold. Based on a 0.50 threshold, the model predicts the correct Sino-graph for 220 out of 315 compounds, an accuracy of 69.8 %. The C-index for the model was 0.693 (Table 1). These metrics of model performance converge in indicating that, although model performance is well above chance, it is also far from perfect. Harrell suggests a C-index of 0.8 as a rough benchmark for predictive models. The informativity-only model falls short of this. A complete account of the corpus on a word-by-word basis requires additional predictors. We return to this issue in the following section. Most of the mistakes made by the model come from predictions based on probabilities in the 0.40–0.60 range (middle of Fig. 5). These are cases for which informativity does not strongly favour one sino-graph over the other. It is in these cases that truncation may be arbitrary or require some other or additional explanation.

We have left investigation of several other possible factors to future work. Compoundhood improved the model enough to earn its keep; headedness did not, although it may still emerge as explanatory in larger corpora. Other structural factors, including grammatical category and argument structure, as well as formal factors, including tone and surface tone sequences, may also account for additional variance. For Chinese, family constituent size might indirectly index semantic factors, i.e., semantically null sino-graphs tend to have large constituent families, but exploring explicit semantic measures of informativity, e.g., semantic specificity, may also prove useful. Raw sino-graph frequency is also an appropriate index of informativity, although it is likely to correlate with family size to some degree. We leave a more complete model of truncation to future work and focus, in the remainder of this paper, on the nature of the informativity-based effects we have established as part of the explanation.

4 The nature of informativity constraints

Truncation patterns in Chinese neologisms have been described as arbitrary. We demonstrated a significant tendency to preserve informative sino-graphs, i.e., those

that minimize uncertainty about the base. Our analysis included four orthogonal predictors based on indices of informativity: family constituent size of the first sino-graph, family constituent size of the second sino-graph, frequency ratio of the first sino-graph and frequency ratio of the second sino-graph. Each of these factors influenced truncation in the direction of preserving the informative sino-graph. Moreover, with the exception of second sino-graph frequency ratio, each factor explained a significant portion of variance in the truncation data.

The non-significant effect of second sino-graph frequency ratio is not surprising from the perspective of lexical processing. Initial material is linguistically prominent and has a privileged role in lexical access (e.g., Becker et al. 2012; see also Walker 2011 for a succinct review). In spoken word recognition, unique initial material, i.e., material that uniquely identifies a word in the lexicon, elicits faster and more accurate responses than unique final material (e.g., Nooteboom 1981). This result suggests that retrospective conditional probabilities are either not deployed in lexical access at all or are relied on much less than prospective conditional probabilities. Our results show that the same is true in truncation. Just as in lexical access, prospective conditional probabilities have a stronger effect than retrospective conditional probabilities.

The fact that frequency ratio is a good predictor of truncation only when it goes in the direction of lexical processing, i.e., forward but not backward conditional probabilities, connects with processing-based theories of morphological productivity. Frequency ratio has also been used, alongside other metrics (e.g., junction strength), to quantify the parseability of a phonological string, the key finding being that parseability correlates with morphological productivity (Hay 2003). Strings that can be easily parsed from speech (low frequency ratios) are likely to be re-used with consistent meanings, i.e., they are morphologically productive. In contrast, strings with high frequency ratios are more likely to be processed together with following material and to be less productive. Frequent holistic processing offers a reasonable explanation for why truncation prefers to spare sino-graphs with high frequency ratios. Namely, sino-graphs with high frequency ratios are likely to evoke the truncated material, facilitating lexical access to the base from the truncated surface form. Models of spoken word recognition that parse prominent material from the signal and infer less prominent material (e.g., Grosjean and Gee 1987) are particularly well-suited to recover the base from truncated surface forms. This kind of processing based-explanation of truncation has cross-linguistic support as well. Alber and Arndt-Lappe (2012:298–299) report that across a corpus of 91 truncation patterns from 27 languages, 50.5 % preserve the left edge of the base, 16.5 % preserve the stressed syllable, 7.7 % the first and stressed syllable, 2.1 % the last syllable, and 8.8 % show other preservation patterns. Truncation patterns tend to preserve material that facilitates lexical activation of the base.

For our other index of informativity, constituent family size, there was a significant effect of both first and second sino-graph family size. Sino-graphs with smaller family sizes make better surface forms regardless of whether they are initial or final in the base. To motivate family size as an index of informativity, we drew on the concept of entropy from information theory. From the standpoint of maximizing system entropy, under-used sino-graphs make better neologisms. This is true regardless of word position. It is worthwhile to note in this context that the constituent families of Chinese

sino-graphs are impressively large. As a basis of comparison, consider English. Plag and Kunter (2010) computed constituent family sizes for English compounds within three large corpora. The majority of constituent family sizes fall between two and 15 members. For example, they report the right constituent family of *office* computed over words contained in a pronunciation dictionary (Teschner and Whitley 2004) as six: *box office*, *head office*, *home office*, *press office*, *ticket office*, *tourist office*. The item *man* was removed from their analysis because it was an extreme outlier, with 71 right constituent family members, e.g. *mailman*, *spiderman*, *henchman*, *etc*. In contrast, the average right constituent family size for Chinese compounds in our corpus was 75. The average left constituent family size was 56. As can be observed in the supplementary material for this paper (and also figures in the Appendix), there are large numbers of sino-graphs with family sizes spanning between 70 and 150 members. This is the case regardless of whether the calculation is over the dictionary or over usage-based corpora such as the Google books corpus (see Appendix for a comparison). Because of the large and varied family sizes in Chinese there is more information at stake in truncation than in languages with smaller constituent families, such as English. Pressure to increase entropy across the lexicon by creating new words from under-used sino-graphs may emerge to a greater degree in a language like Chinese with comparatively large constituent families than in a language like English, with smaller family sizes.

Chinese offers a rich testing ground for the role of informativity in truncation. In this study, we focused on neologisms listed in the dictionary, but the truncation patterns in these words are part of a broader phenomenon involving lengthening and shortening of words according syntactic, semantic and prosodic factors (Duanmu 2012). Duanmu (2013) estimates that 90 % of Chinese words have "elastic length" in that they alternate between monosyllabic and disyllabic forms according to context. For example, 商店 *shāngdiàn* 'business-store' is shortened to 店 *diàn* 'store' in certain linguistic contexts. In noun-noun polysyllabic compounds such as 煤炭+商店 *méitàn* 'coal-coal' + *shāngdiàn* 'business-shop', truncation tends to apply to the second compound, so that 煤炭+商店 becomes 煤炭+**店** *méitàn-diàn* 'coal-coal-store' instead of the first compound, *煤商店 *méi-shāngdiàn* 'coal-business-store'. There is substantial scope for developing more refined models of truncation integrating informativity with grammatical triggers.

We can imagine several ways in which informativity might be integrated with other factors influencing truncation patterns. A key question in considering these options is whether informativity should be treated as a macro-constraint or broken down into different components. Informativity effects could potentially be derived from the interaction of simple constraints, which may relate more directly to the individual factors that went into our regression model as opposed to the truncation predictions that came out of it. Alternatively, informativity may have a complex computation but play a simple role in the grammar as is the case, for example, in the implementation of frequency effects in Noisy Harmonic Grammar (Coetzee and Kawahara 2013).

The truncation probabilities predicted by our logistic regression model (Fig. 5) are based on four orthogonal indices of informativity. When the four indices influence outcomes in the same direction, then the model generates high predicted probabili-

ties. An example is 撤出 *chèchū* 'remove', which is truncated to 撤 *chè* (not 出 *chū*) in the neologism 撤资 *chèzī* 'remove funds'. The first sino-graph, 撤 *chè* 'remove', has a high frequency ratio, 0.19, and a small family size, 29 members; the second sino-graph, 出 *chū* 'go out', has a low frequency ratio, 0.01, and large family size, 426 members. All four informativity indices point to 撤 *chè* 'remove' as the most informative truncation of 撤出 *chèchū* 'remove'. The model predicts the probability of truncating 出 *chū* (and preserving 撤 *chè*) to be 0.96. In other cases, indices of informativity conflict. A sino-graph in a particular word may be comparably informative on frequency ratio (high frequency ratio) but uninformative on family size. In these cases, truncation probabilities hover around 0.50. If informativity is a macro-constraint to which each index of informativity may contribute, then high predicted probabilities should correlate with model accuracy. In other words, the informativity-based model should do best when all indices of informativity point in the same direction.

To evaluate this, we re-examined the predicted probabilities generated according to Eq. (4). As shown in Fig. 5, the distribution of predicted probabilities is wide. The compounds predicted to follow the dominant pattern of first sino-graph preservation have probabilities that span the entire possible range, from 0.5 to 1. When the decision threshold for determining a categorical prediction is set to 0.50, probabilities ranging widely from 0 to 0.50 (first sino-graph) and from 0.50 to 1 (second sino-graph) are all treated as making the same prediction. Instead of flattening the variation expressed in Fig. 5 into categorical predictions, we now make use of it to assess whether the data are patterning according to a macro-constraint on informativity.

To bring out the gradience inherent in the predicted probabilities, we scaled the decision threshold for interpreting probabilities in terms of categorical predictions from 0.50 to 0.95 in 0.05 steps. Recall that based on a 0.50 threshold the model predicts the correct Sino-graph in 69.8 % of the cases. As the confidence threshold is increased, model accuracy also increases. For example, at a decision threshold of 0.65, we relieve the model from having to predict the 71 compounds with low confidence (probabilities between 0.35 and 0.65). These are cases in which informativity does not make clear predictions, either because of small differences between competing sino-graphs or because of conflicting informativity indices. On the remaining 244 compounds, the model makes the correct prediction for 216 compounds (28 wrong), an accuracy rate of 88.5 %. Further increasing the threshold yields still better performance. At a threshold of 0.80, model accuracy improves to 93.3 %. At such a high threshold, the model is making decisions on 75 compounds. The gradient relationship between model accuracy and decision threshold is shown in Fig. 6. Monotonic increases in model accuracy result from increases in decision threshold. This is precisely the behaviour expected of an informativity-based macro-parameter. As differences in informativity between sino-graph candidate increases, model accuracy also increases. This indicates that, when it comes to formalizing informativity effects, an aggregate measure, drawing on several orthogonal indices of informativity, is appropriate.

The result that model accuracy scales with decision threshold speaks to the validity of an informativity-based macro-constraint on truncation. For some words, both

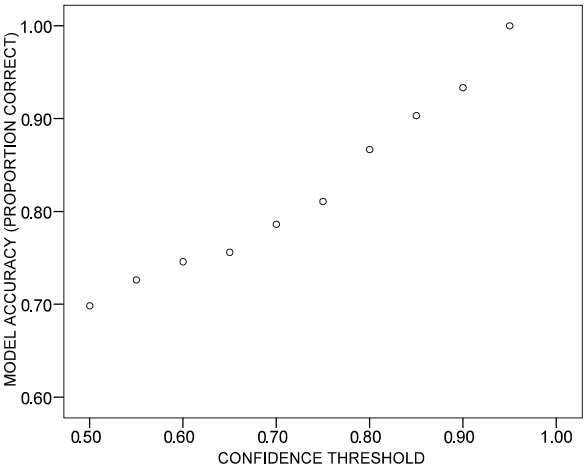

Fig. 6 Model accuracy, expressed as the proportion of correct predictions, as a function of model confidence

sino-graphs are similarly informative. These are cases for which predicted probabilities are near 0.50. When forced to make predictions about these marginal cases, the model performs at well above chance, nearly 70 % accuracy. When one member of the base is clearly more informative than the other, i.e., predicted probabilities farther from 0.50, model accuracy surges. Informativity has the greatest influence on the form of truncations just when uncertainty is at stake. This suggests that there may be an upper bound on information loss, at least in cases such as truncation where meaning is conveyed in part by the relationship between surface forms. Word-formation processes that appear to flout morphological systematicity of the conventional sort (see, e.g., Haspelmath and Sims 2013:50 on word formation operations that fall outside the domain of morphology) may be those that approach minimum information requirements. In particular, the range of possible truncations may be limited to those that preserve sufficient information about the base.

5 Conclusion

The information content of sino-graphs, as encapsulated in measures of family size and frequency ratio, influences patterns of truncation in Chinese neologisms. An informativity-based logistic regression model predicted both the dominant trend in the data—a tendency for the first sino-graph to surface—as well as "exceptions", the still large number of cases in which the second sino-graph surfaces. The sino-graphs that survive truncation tend to be those that contain the most information about the underlying word. As the information gap between candidate outputs increases, the tendency to select the informative candidate strengthens. We interpret this result as evidence for a constraint regulating the degree of uncertainty between base and truncation.

Acknowledgements We would like to thank San Duanmu for comments and discussion about several aspects of this paper, particularly Chinese pseudo-compounds, Mike Opper for comments on an earlier

version, Cathi Best for discussion, and audiences at the 20th Annual Conference of the International Association of Chinese Linguistics (IACL) at Hong Kong Polytechnic University and the conference on Morphology and Its Interfaces at the University of Lille 3 for their feedback. For general discussion about information theory and phonology, we thank Beth Hume, Andy Wedel and the Sydney Phonology Reading group, especially Ivan Yuen, Katherine Demuth, and Mark Harvey. Remaining shortcomings are solely the responsibility of the authors. Work on this project was partly supported by an Australian Research Council grant (DE120101289) to the first author.

Appendix: Relationship between informativity measurements in different corpora

Figures 7 and 8 plot family size calculated from the Google corpus of Chinese books, x-axis, against family size calculated from our compiled lexicon (Figs. 1 and 2). Markers for family sizes were set according to whether the sino-graph was truncated in the abbreviation or whether it surfaced in the abbreviation. Circles indicate compounds for which the first sino-graph surfaced in abbreviations and squares indicate compounds for which the second sino-graph surfaced. A regression line was fit to all of the markers. The position of the markers, circles and squares, echo the trends revealed in pyramid plots in Figs. 1 and 2 in Sect. 2. The concentration of circles (relative to squares) is greatest in the lower-left corner of Fig. 7, which shows the family size of the first sino-graph. This indicates that family sizes tend to be smaller for sino-graphs that surface in neologisms than for sino-graphs that are truncated. In Fig. 8, which shows the family size of the second sino-graph, the density of squares is greatest in the lower-left corner of the figure. For both the first sino-graph of compounds, Fig. 7, and the second sino-graph of compounds, Fig. 8, family sizes tend to be smaller for sino-graphs that surface in abbreviations than for sino-graphs that are truncated.

Although the family size calculations are similar across the dictionary and the book corpus, there are also some notable differences. These can be identified with reference to the regression line. Most of the large deviations are located above the regression line. These deviations are cases in which the dictionary provides larger family sizes than the Google book corpus, and they may have implications for how well family constituent size predicts truncation.

Figure 9 compares frequency ratios calculated in the Gigaword corpus with those calculated in the Google books corpus. Frequency ratios calculated over the Google corpus, x-axis, are plotted against frequency ratios values calculated in the Gigaword corpus, y-axis. The circles show the frequency ratio for sino-graphs that survive truncation to surface in neologisms. The squares show the frequency ratio for sino-graphs that were truncated. The general trend with respect to circles and squares echoes what was displayed in the pyramid plots Figs. 3 and 4. The density of circles relative to squares increases with frequency ratio. This indicates that sino-graphs with higher frequency ratios are more likely to survive truncation. A linear regression line fit to all of the values is shown in solid black. The number and size of deviations on both sides of the regression line (above and below) is fairly balanced and contain roughly equal numbers of circles and squares. This suggests that the two corpora return frequency ratios that are highly comparable or, at least, that there are not systematic differences in frequency ratio calculations due to the nature of these corpora.

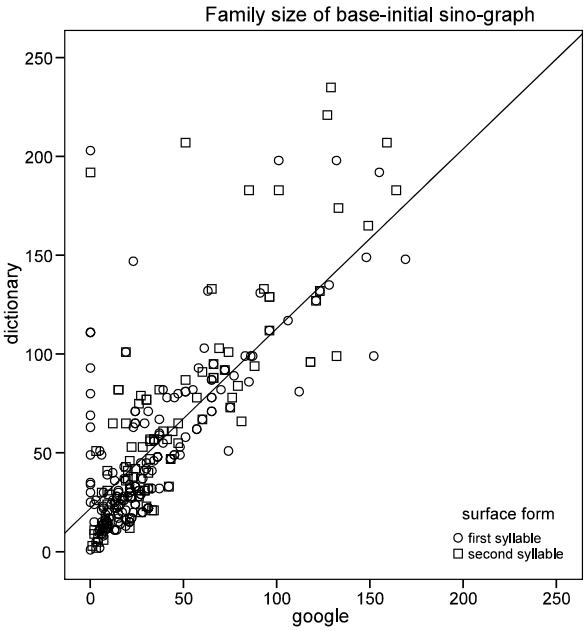

Fig. 7 Scatter plot of first (base-initial) sino-graph family size calculated from the Google book corpus, x-axis, against first sino-graph family size calculated from the dictionary. The *solid line* shows a linear regression line fit to all of the data. The *circles* show the family size for sino-graphs that surface in neologisms. The *squares* show the family size of sino-graphs that are truncated

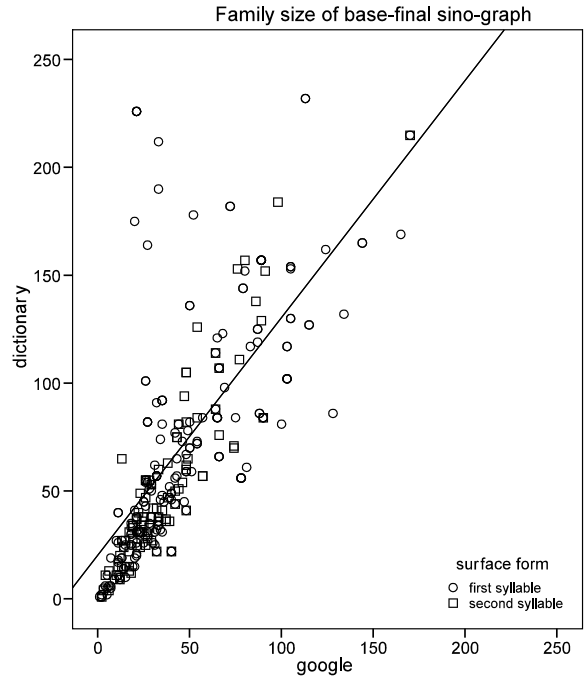

Fig. 8 Scatter plot of second (base-final) sino-graph family size calculated from the Google book corpus, x-axis, against second sino-graph family size calculated from the dictionary. The *solid line* shows a linear regression line fit to all of the data. The *squares* show the family size of sino-graphs that are truncated in neologisms. The *circles* show the family size of sino-graphs that surface

Figure 10 plots frequency ratios for the second (base-final) sino-graph calculated across the two corpora. The format follows that of Fig. 9. The x-axis shows the frequency ratios calculated from google and the y-axis shows the ratios calculated from Gigaword.

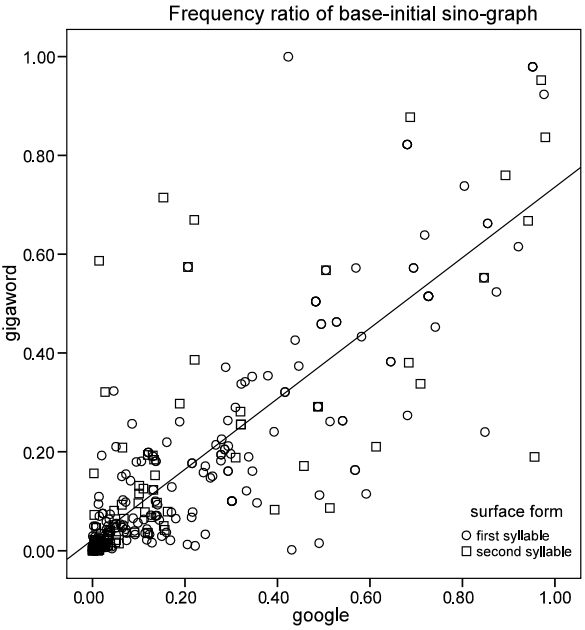

Fig. 9 Scatter plot of first (base-initial) sino-graph frequency ratio calculated from the google book corpus, *x*-axis, against first sino-graph frequency ratio calculated from a sub-section of the Gigaword corpus. The *solid line* shows a linear regression line fit to all of the data. The *squares* show the frequency ratios of sino-graphs that are truncated in abbreviations. The *circles* show frequency ratios of sino-graphs that surface in the abbreviations

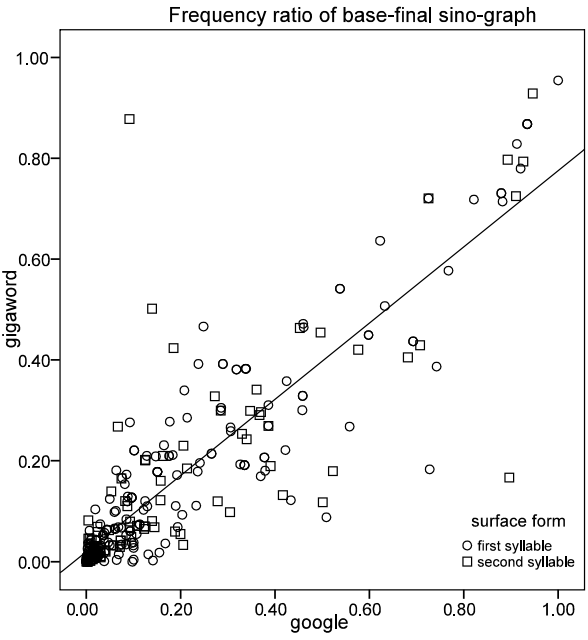

Fig. 10 Scatter plot of second (base-final) sino-graph frequency ratio calculated from the google book corpus, *x*-axis, against second sino-graph frequency ratio calculated from a sub-section of the Gigaword corpus. The *solid line* shows a linear regression line fit to all of the data. The *squares* show the frequency ratios of sino-graphs that are truncated in abbreviations. The *circles* show frequency ratios of sino-graphs that surface in the abbreviations

References

Alber, B., & Arndt-Lappe, S. (2012). Templatic and subtractive truncation. In *The phonology and morphology of exponence-the state of the art. in print*, Oxford: OUP.

Albright, A. (2002). Islands of reliability for regular morphology: evidence from Italian. *Language*, *78*(4), 684–709.
Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, *90*, 119–161.
Aronoff, M. (1994). *Morphology by itself*. Cambridge: MIT Press.
Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3059.
Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
Becker, M., Nevins, A., & Levine, J. (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language*, *88*(2), 231–268.
Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.
Bell, M., & Plag, I. (2012). Informativeness is a determinant of compound stress in English. *Journal of Linguistics*, *48*(3), 485–520.
Bell, M., & Plag, I. (2013). Informativity and analogy in English compound stress. *Word Structure*, *6*(2), 129–155.
Benua, L. (1995). Identity effects in morphological truncation. In J. Beckman, L. Walsh Dickey, & S. Urbanczyk (Eds.), *University of Massachusetts occasional papers: Vol. 18. Papers in optimality theory* (pp. 77–136). Amherst: GLSA Publications.
Benua, L. (1997). *Transderivational identity: phonological relations between words*. Ph.D. Dissertation, University of Massachusetts, Amherst.
Blevins, J. P. (2013). The information-theoretic turn. *Psihologija*, *46*(4), 355–375.
Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.
Burzio, L. (2002). Surface-to-surface morphology: when your representations turn into constraints. In *Many morphologies* (pp. 142–177).
Calhoun, S. (2010). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, *25*(7–9), 1099–1140.
Ceccagno, A., & Basciano, B. (2007). Compound headedness in Chinese: an analysis of neologisms. *Morphology*, *17*(2), 207–231.
Coetzee, A. W., & Kawahara, S. (2013). Frequency biases in phonological variation. *Natural Language & Linguistic Theory*, *31*(1), 47–89.
Duanmu, S. (1999). Stress and the development of disyllabic words in Chinese. *Diachronica*, *16*(1), 1–35.
Duanmu, S. (2012). Word-length preferences in Chinese: a corpus study. *Journal of East Asian Linguistics*, *21*(1), 89–114.
Duanmu, S. (2013). How many Chinese words have elastic length. In *Eastward flows the Great river: Festschrift in honor of Prof. William S.-Y. Wang on his 80th birthday* (pp. 1–14). Hong Kong: City University of Hong Kong
Graff, D., & Chen, K. (2005). Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN, 1*, 58563-58230.
Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, *25*(1), 135–155.
Hall, K. C. (2012). Phonological relationships: a probabilistic model. *McGill Working Papers in Linguistics*, *22*(1), 1–14.
Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Berlin: Springer.
Haspelmath, M., & Sims, A. (2013). *Understanding morphology*. London: Routledge.
Hay, J. (2003). *Causes and consequences of word structure*. London: Routledge.
Hay, J., & Baayen, H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, *9*(7), 342–348.
Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, *63*(1), 305–340.
Hume, E., Hall, K. C., Wedel, A., Ussishkin, A., Adda-Decker, M., & Gendrot, C. (2013). Anti-markedness patterns in French epenthesis: an information-theoretic approach. In *Proceedings of the annual meeting of the Berkeley linguistics society* (Vol. 37).
Hume, E., & Mailhot, F. (2013). The role of entropy and surprisal in phonologization and language change. In A. Yu (Ed.), *Origins of sound change: approaches to phonologization*, Oxford: Oxford University Press.

Kenstowicz, M. (1996). Base-identity and uniform exponence: alternatives to cyclicity. In J. Durand & B. Laks (Eds.), *Current trends in phonology: models and methods* (pp. 363–393). Paris-X and Salford: University of Salford Publications.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations* (pp. 169–174). Association for computational linguistics.

McCarthy, J. J., & Prince, A. (1995a). Faithfulness and reduplicative identity. In J. Beckman, L. Walsh Dickey, & S. Urbanczyk (Eds.), *University of Massachusetts occasional papers in linguistics* (Vol. 18, pp. 249–384). Amherst: GLSA Publications.

McCarthy, J. J., & Prince, A. (1995b). Prosodic morphology. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 318–366). Cambridge/Oxford: Blackwell.

McCarthy, J. J., & Prince, A. (1999). Faithfulness and identity in prosodic morphology. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The prosody-morphology interface* (pp. 218–309). Cambridge: Cambridge University Press.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: an information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In *Analogy in grammar: form and acquisition* (pp. 214–252).

Nagano, A., & Shimada, M. (2014). Morphological theory and orthography: Kanji as a representation of lexemes. *Journal of Linguistics*, *50*(02), 323–364.

Nooteboom, S. G. (1981). Lexical retrieval from fragments of spoken words: beginnings vs. endings. *Journal of Phonetics*, *9*(4), 407–424.

Opper, M., & Sugar, A. (2012). *Truncation and headedness in Chinese compounding: a dictionary-based study*. Paper presented at the 25th annual North American conference on Chinese linguistics, Ann Arbor, MI.

Packard, J. (2000). *The morphology of Chinese: a linguistic and cognitive approach*. Cambridge: Cambridge University Press.

Pan, S., & Hirschberg, J. (2000). Modeling local context for pitch accent prediction. In *Proceedings of the 38th annual meeting on association for computational linguistics* (pp. 233–240). Association for computational linguistics.

Plag, I., & Baayen, H. (2009). Suffix ordering and morphological processing. *Language*, *85*(1), 109–152.

Plag, I., & Kunter, G. (2010). Constituent family size and compound stress assignment in English. *Linguistische Berichte, Sonderheft 17*.

Priva, U. C. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th west coast conference on formal linguistics* (pp. 90–98).

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

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

Sproat, R., & Shih, C. (2002). Corpus-based methods in Chinese morphology. In *Tutorial at the 19th COLING*.

Teschner, R. V., & Whitley, M. S. (2004). *Pronouncing English: a stress-based approach, with CD-rom*. Georgetown University Press.

Tily, H., & Kuperman, V. (2012). Rational phonological lengthening in spoken Dutch. *The Journal of the Acoustical Society of America*, *132*(6), 3935–3940.

Walker, R. (2011). *Vowel patterns in language* (Vol. 130). Cambridge: Cambridge University Press.

Wedel, A., Jackson, S., & Kaplan, A. (2013a). Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, *56*(3), 395–417.

Wedel, A., Kaplan, A., & Jackson, S. (2013b). High functional load inhibits phonological contrast loss: a corpus study. *Cognition*, *128*(2), 179–186.

Yuan, L. (2002). *The contemporary Chinese dictionary*. Foreign Language Teaching and Research Press.