Productivity and semantic transparency: An exploration of word formation in Mandarin Chinese

Shen Tian[1] & Harald Baayen[2]
1. Northwest University, Xi'an
2. Eberhard-Karls University, Tübingen

Abstract

We used word embeddings to study the relation between productivity and semantic transparency. We compiled a dataset with around 2700 two-syllable compounds that shared position-specific constituents (henceforth pivots), and some 1100 suffixed words. For each pivot and suffix, we calculated measures of productivity as well as measures of semantic transparency. For compounds, productivity ($\mathcal{P}$) was negatively correlated with the number of types (V) and with the semantic similarity between non-pivot constituents and their compounds. Conversely, greater semantic similarity of the pivot with either the compound or the non-pivot constituent predicted higher degrees of productivity. Visualization with t-SNE revealed clustering of suffixed words' embeddings, but no by-pivot clustering for compounds, except for a minority of pivots whose regions in semantic space did not contain intruding unrelated compounds. A subset of these pivots were found to realize a fixed shift in semantic space from base word to compound, a property that also emerged for several suffixes. For these pivots, no correlation between $\mathcal{P}$ and V was present. Thus, Mandarin compounds appear to realize, at one extreme, motivated but unsystematic concept formation (where other pivots could just as well have been used), and at the other extreme, systematic suffix-like semantics.

1 Introduction

Aronoff (1976) characterized morphological productivity as a central mystery in word-formation. In this study, we address the mystery of the productivity of compounding in Mandarin Chinese. Most studies of morphological productivity have focused on derivational word formation. Because the relation between form and meaning is much less clear for compounds compared to derived words, structuralist analyses have argued that compounds do not form morphological categories, and are not rule-governed (Schultink 1961; Aronoff 1976; Booij 1977; Bauer 1983; Marle 1985). Nevertheless, compounding can be productive. In fact, in Mandarin Chinese, compounding is the main word formation process. Compounds account for 70-80% of all words (Institute of Language Teaching and Research). For neologisms, compounds account for around 95% of word types (Ceccagno and Basciano 2007).

In this study, we report a series of investigations into the productivity of Mandarin compounding, with the aim of improving our understanding of under what conditions a Mandarin word (e.g., 大, *da4*, 'big') can be productive as a position-specific constituent in Mandarin compounds (e.g., 大家, *da4jia1*, 'everyone'). In what follows, we use the term 'pivot' to refer to position-specific constituents in compounds that form series (e.g., 大家, *da4jia1*, 'everyone'; 大学, *da4xue2*, 'university'; 大量, *da4liang4*, 'generous'). We refer to these series as pivot (morphological) families. Pivots in compounds are formally similar to affixes in derived words, with crucial difference that affixes tend not to occur by themselves as independent words.

Given a pivot and its compound family, the question arises of how semantically transparent the compounds with a given pivot are, and how semantic transparency co-varies with productivity. Shen and Baayen (2021) addressed this question for adjective-noun compounds, using distributional semantics to assess semantic similarity. In section 2, we introduce some key findings, building on their data, which we extended with verbal and nominal pivots. Section 3 zooms in on the geometry of pivot families in distributional semantic space. We present evidence that in Mandarin Chinese, the productivity of a given target pivot decreases when words formed with other pivots intrude into the semantic cloud of the target pivot family, and hence are too close in meaning to the words of the target pivot family itself. Conversely, pivots with family clouds without intruders show more consistent semantic relations between pivot words and the compounds in the pivot family. Section 4 extends this line of research by examining the geometrical properties of Mandarin derivational suffixes, which provide stronger evidence for clustering and for more systematic relations with their base words.

In the general discussion, we reflect on the implications of our findings for our understanding of the different ways in which Mandarin implements word formation, and the concomitant consequences for morphological productivity.

2 Productivity and semantic transparency

2.1 Data

The compounds under investigation in the present study contained one of three different kinds of pivots, varying from antonymous senses to synonymous senses, from adjectival pivots to verbal and nominal pivots, and with pivots at the initial position as well as the second position in a two-character compounds. First, 28 pairs of high-frequency gradient adjectives (e.g., hot vs. cold; big vs. small) were selected. Next, all compounds sharing one of these pivots, and that had a noun as second constituent, were extracted from the Chinese National Corpus (http://corpus.zhonghuayuwen.org/). 大家 (*da4jia1*, 'everyone'), 大学 (*da4xue2*, 'university') and 大量 (*da4liang4*, 'generous') are examples of compounds with the pivot 大 (*da4*, 'big').

Second, 14 pairs of verbal antonyms (go up/go down, remember/forget, buy/sell, etc.) were selected and all compounds with these verbs in initial position were extracted from the Chinese National Corpus. Examples of compounds with verbal pivots are 上学 (*shang4xue2*, go up school, 'go to school'), 记仇 (*ji4chou2*, remember hatred, 'hold a grudge'), and 卖国 (*mai4guo2*, sell country, 'treason').

The final set of compounds was based on a series of roughly synonymous nominal pivots as second constituent. These 19 nominal pivots were restricted to words denoting people, resulting in compounds such as 高徒 (*gao1tu2*, tall apprentice, 'brilliant student'), 影迷 (*ying3mi2*, 'movie fan') and 毒枭 (*du2xiao1*, drug man, 'drug pusher'). Table 1 provides an overview of these compound types and their constituents. Though the selection of pivots is designed to be diverse, the pivots that we sampled necessarily represent only a small subset of adjectives, verb, and nouns that are found in Mandarin compounds.

On top of this, we selected 1093 two-character suffixed words with the 7 most common Mandarin suffixes, namely, -子 (*zi4*, 'diminutive marker'), -者 (*zhe3*, 'man'), -化 (*hua4*, '-ise'), -头 (*tou2*, noun marker), -们 (*men2*, 'plural marker'), -家 (*jia1*, 'certain kind of people'), and 儿 (*er2*, 'diminutive marker'), in order to allow comparisons between compounding and suffixation.

Table 1: Properties two-character Mandarin compounds with different pivots. For Adjective-Noun compounds (AN), the gradable adjectival pivots are located at the initial position of a two-character compound with a non-pivotal noun. For Verb-Noun compounds (VN), the antonymous verbs occupy the first position of a two-character compound with a non-pivotal noun. XN represents two-character compounds with synonymous nominal pivots at the second position; the non-pivots X in an XN formation are not restricted with respect to part of speech. It is worth noting that a few compounds in our sample contain two pivots, such as 爱人 (*ai4ren2*, 'lover'), the first constituent of which is one of our selected verbal pivots, and the second constituent of which is one of our selected nominal pivots.

Compound type	**AN**	**VN**	**XN**
Constituent configuration	gradable adjective + noun	antonymous verb + noun	any POS + synonymous noun
Structure	[**pivot**] [non-pivot]	[**pivot**][non-pivot]	[non-pivot][**pivot**]
Example	大家 (*da4jia1*, big family, 'everyone')	上学 (*shang4xue2*, up school, 'go to school')	毒枭 (*du2xiao1*, drug man, 'drug pusher')
Number of pivots	56	28	19
Number of compounds	1440	467	808

2.2 Productivity

Several quantitative measures are available for gauging morphological productivity. Type counts $V(N)$ based on N tokens capture one aspect of what Corbin (1987) and Bauer (2001) refer to as "profitability", the extent to which a word formation process has already been used. However, it only presents what words are in use without providing insight into the probability of coinages in the future. To assess productivity with respect to the morphological category itself, we can calculate the probability $\mathcal{P}(\text{piv}, N_{\text{piv}})$ that a token containing a pivot constituent is a two-character Mandarin compound that has not been seen before given that we have seen N_{piv} tokens of this type of compound:

$$\mathcal{P}(\text{piv}, N_{\text{piv}}) = \frac{V(1, \text{piv}, N_{\text{piv}})}{N_{\text{piv}}},$$

where $V(1, \text{piv}, N_{\text{piv}})$ refers to the number of hapax legomena, the words with frequency equal to one, and N_{piv} denotes the number of tokens with the pivot. In what follows, we will mostly use a simplified notation in which the dependency of these measures on the sample size N is not specified.

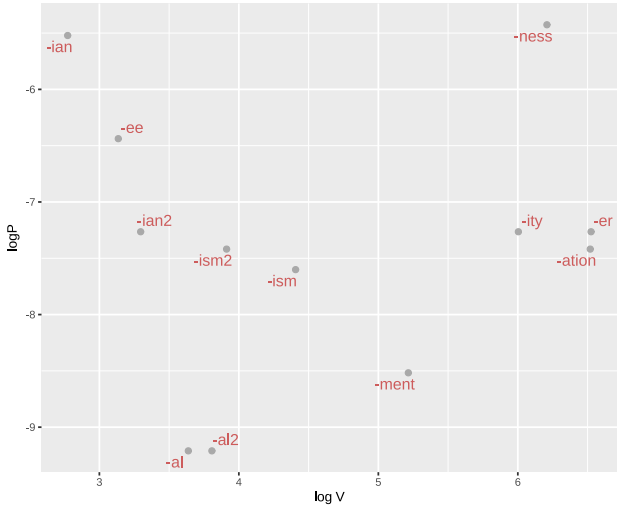

Figure 1: Type counts V and productivity $\mathcal{P}$ for the noun-forming English suffixes studied by Baayen & Lieber (1991). *-ian*, *-ism* are deadjectival nominalizing affixes, *-ian2*, and *-ism2* denote the denominal nominalizing affixes. *-al* represents the deverbal nominalizing suffix, and *-al2* represents its denominal counterpart.

Figure 1 presents 12 nominalizing English affixes in the plane defined by $\log V$ and $\log P$, using the counts reported in (Baayen and Lieber 1991). There is no clear relation between the two measures. However, as shown in Figure 2, a similar plot for Mandarin pivots reveals strong evidence for a negative and linear functional relation between $\log \mathcal{P}$ and $\log V$. For instance, the noun 人 (*ren3*, 'man'), which is characterized by a large number of types, shows up with a rather small value of $\log \mathcal{P}$. Conversely, 甜 (tian2, sweet) has a low type count, but a higher productivity close to that predicted by the regression line.

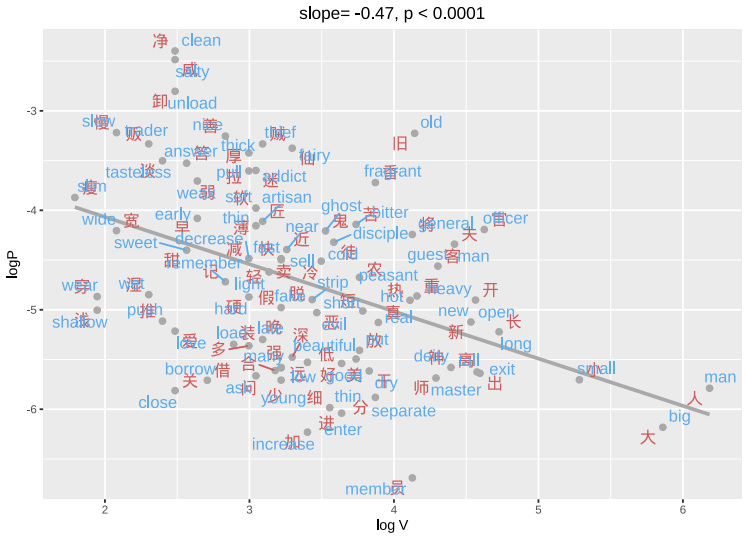

Figure 2: Scatterplot of pivots in the plane spanned by log V and log $\mathcal{P}$. The grey line represents the least squares linear regression line. As log V increases, log $\mathcal{P}$ tends to decrease.

Shen and Baayen (2021) observed a similar negative correlation for the subset of compounds with an adjective as pivot. They hypothesized that adjectival pivots are more likely to have similar semantics compared to English derivational affixes, which realize a wide range of different semantic functions. However, the present study incorporates a broader range of pivots that realize more diverse semantics. For instance, verbs are used in the common 'predicate+object' construction in which the second constituent realizes the direct object (e.g, 借钱, *jie4qian2*, 'borrow money'). They are also found in less common 'predicate+object' construction, in which the noun constituent denotes manner, result, purpose, or instrument (e.g., 出丑, *chu1chou3*, go out ugly, 'make a fool of somebody'). In our dataset, pivot nouns (constrained to the second position in our dataset), can combine with adjectives in an attributive construction (i.e., 良师, *liang2shi1*, good teacher, 'mentor'), with nouns in a subordinate construction (i.e., 画师, *hua4shi1*, draw master, 'painter'), in a coordinate construction (鬼神, *gui3shen2*, ghost god, 'supernatural being'), and with verbs in a 'predicate-object' structure (i.e., 敬神, *jing4shen2*, 'pray to a god'). Furthermore, the semantic relation of the constituents in X+N compounds also varies substantially (compare, e.g., 铁匠, *tie3jiang4*, iron smith, 'blacksmith' and 书迷, *shu1mi2*, book enthusiast, 'bookworm'). Since the semantics realized by the broad range of pivots in the present study are highly variable, and since we nevertheless still observe a negative linear relation between log $\mathcal{P}$ and log V, we are forced to conclude that the explanation offered by Shen and Baayen (2021) for this linear relation is incorrect.

As a first step towards obtaining an improved understanding of the relation between category-conditioned productivity $\mathcal{P}$ and extent of use V, in the next section we consider the relation between semantic transparency and productivity.

2.3 Semantic transparency

Following Shen and Baayen (2021), we made use of three measures gauging the semantic transparency of pivot families. For each compound in our dataset, we extracted its pre-calculated embedding from the Tencent resource at https://ai.tencent.com/ailab/nlp/en/embedding.html, and used these embeddings to calculate the following measures: [1]

1. $s_{\text{piv-com}}$: the correlation, averaged over all compounds with a given pivot, of the embedding of the pivot and the embeddings the compounds;

2. $s_{\text{nonpiv-com}}$: the correlation, averaged over all compounds with a given pivot, of the embedding of the non-pivotal constituent and the embedding of the compound;

3. $s_{\text{piv-nonpiv}}$: the correlation, averaged over all compounds with a given pivot, of the embedding of the pivot and the embedding of the non-pivotal constituent.

Regression modeling revealed good evidence for a positive correlation between $s_{\text{piv-com}}$ and $\mathcal{P}$ ($t(82) = 4.58, p < 0.0001$), as well as between $s_{\text{piv-nonpiv}}$ and $\mathcal{P}$ ($t(82) = 10.05, p = 0.006$), see Figures 3 and 4. For $s_{\text{nonpiv-com}}$ and $\mathcal{P}$, however, a negative correlation is perhaps present ($t(82) = -2.14, p = 0.04$), see figure 5. In other words, the productivity of a pivot, as assessed with $\mathcal{P}$, increases with the two measures of semantic transparency that compare the embedding of the pivot with the embeddings of either the non-pivot constituent, or the carrier compound. These positive correlations make sense: the more predictable the semantic contribution of a pivot is, the more regular it is, and the more available it should be for the creation of interpretable and understandable novel compounds.

The possibility of a negative correlation for $s_{\text{nonpiv-com}}$ and $\mathcal{P}$ suggests that possibly the transparency of the non-pivot constituent might be interfering with the transparency of the pivot. After all, the non-pivot is a non-pivot only because in our dataset, it is not included as a pivot. However, non-pivots are themselves the pivots of other pivot families. This negative correlation is reminiscent of an observation for English compounds made by Tarasova (2013), who reported that in English a constituent is more productive either as head, or as modifier, but not in both positions, and that more productive constituents are more likely to realise a single semantic relation in their constituent families. In what follows, we therefore investigate in more detail the geometry of the pivot families in the semantic space of Mandarin, with as working hypothesis that it is detrimental for the productivity of a pivot if its cloud of family members in semantic space overlaps substantially with the cloud of family members of other pivots.

[1] One drawback of the semantic vectors that we used is that different senses of the same orthographic word are not distinguished. For instance, 分子 (*fen1zi3*) can mean 'molecule' but also, with different tones, *fen4zi3* meaning 'member'. In the present study, we manually selected the senses with the highest frequency for all homographs with senses that have different pronunciations.

Figure 3: Scatterplot of category-conditioned productivity log$\mathcal{P}$ as a function of $s_{\text{piv-com}}$. The more semantically transparent the pivot is with respect to the meaning of its carrier compounds, the **more** productive the pivot is.

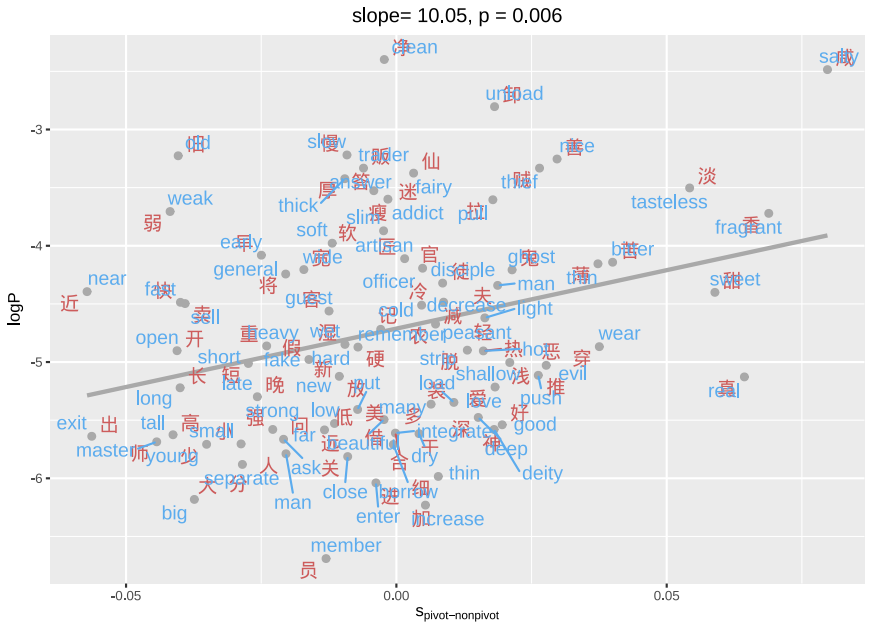

Figure 4: Scatterplot of category-conditioned productivity log$\mathcal{P}$ as a function of $s_{\text{piv-nonpiv}}$. The more similar the embedding of the pivot is to the embedding of the non-pivot constituent, the **more** productive the pivot is.

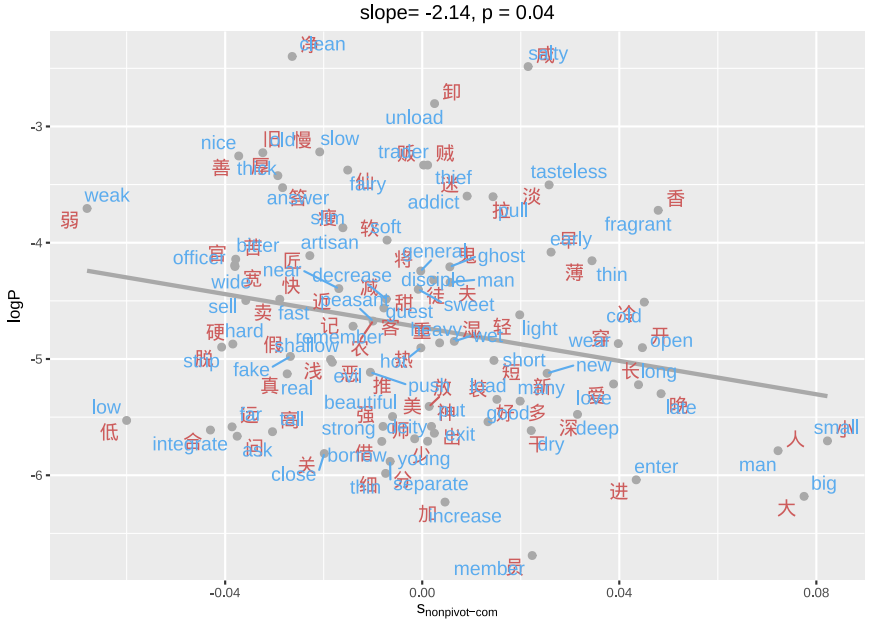

Figure 5: Scatterplot of category-conditioned productivity log$\mathcal{P}$ as a function of $s_{\text{nonpiv-com}}$. The more semantically transparent the non-pivot constituent is with respect to the meaning of its carrier compound, the **less** productive the pivot is.

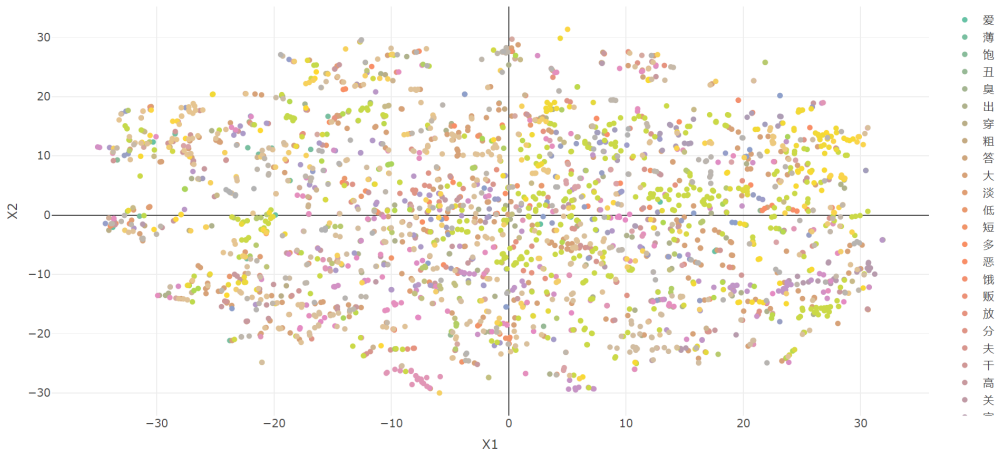

Figure 6: Compounds, color-coded for 82 pivot constituents in a t-SNE plane. No well-defined clusters are present: for any pivot, family members are widely spread out in the 2D map (interactive plot here).

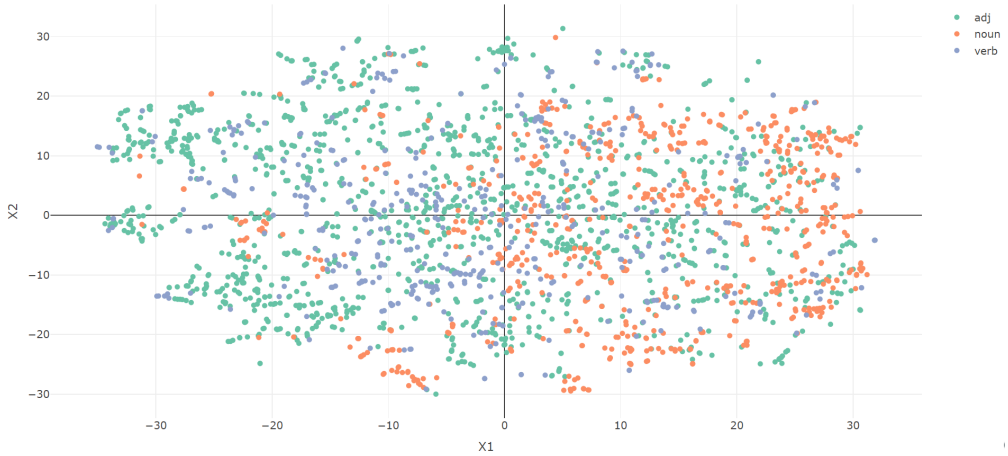

Figure 7: Compounds with 82 pivot constituents in t-SNE space, color-coded by the word category of pivots. Compounds with noun pivots occur more to the right (in orange). Compounds with adjective (green) and verb (blue) pivots are widely dispersed (interactive plot here).

3 The semantic geometry of Mandarin compounds

In order to better understand the geometry of the semantic space of Mandarin, we made use of t-distributed stochastic neighborhood embedding, henceforth t-SNE (Van der Maaten and Hinton 2008). Surprisingly, unlike for English (Shafaei-Bajestan et al., this volume), Russian (Chuang et al., this volume) and Finnish (Nikolaev et al, this volume) nominal inflection, a t-SNE plot does not reveal any well-delineated clusters for Mandarin compounds, as can be seen in Figure 6, similar to what Stupak & Baayen (this volume) report for German particle verbs. When the plot is color-coded for the word category of the pivot, we can observe a tendency for nouns to occur in the right half of the plot, but adjectives and verbs remain widely dispersed (see Figure 7).

Why is it that t-SNE does not detect clusters for Mandarin pivots? To address this question, we reasoned that if a pivot family forms a cluster in semantic space that is well-separated from the clusters of other pivots, the t-SNE algorithm should be able to detect it. However, if clusters of pivot families overlap to a considerable extent, then these clusters should be invisible to t-SNE, as effectively no distinct clusters would actually be present.

In order to investigate this possibility, we first calculated for each pivot its mean vector, i.e., the vector obtained by calculating the average of the embeddings of all its pivot family members. This mean vector represents the centroid of the cluster of the pivot family. For instance, the centroid for 爱 (*ai4*, 'love') is represented by the mean of all the semantic vectors that have 爱 as the first constituent.

As a next step, we calculated the correlations of 爱's pivot family members with this centroid vector, and registered both the mean and the standard deviation of this distribution of correlations. This makes it possible to calculate a 95% confidence interval of this pivot.

After having obtained mean and standard deviation for all 82 pivot clusters, we inspected which pivots have no compounds from other pivot clusters within their 95% confidence interval. In what follows, we refer to pivots with this property as "intruder-free" pivots. Figure 8 presents the compounds with one of the 11 intruder-free pivots with more than 10 family members in the t-SNE 2D plane. Some clustering structure now emerges.

This result raises the question of whether the productivity of a pivot can be predicted from the number of intruders n_I that originate from other pivots. A regression model with $\mathcal{P}$ as response variable and n_I as predictor provided some support for a negative correlation ($\hat{\beta} = -0.2, p = 0.01$). As expected, the more semantic intruders a pivot constituent has, the less productive it may be.

We obtained stronger evidence for a correlation between semantic transparency and the number of intruders. When we gauge semantic transparency with the average of the correlations of the pivot's embedding and the embeddings of its compound family members (i.e., $s_{\text{piv-com}}$), we find a negative correlation with the number of intruders n_I ($\hat{\beta} = -6, 6, p = 0.0006$). Conversely, a linear model regressing the semantic transparency between non-pivot and the carrier compounds (i.e., $s_{\text{nonpiv-com}}$) on n_I reveals a regression line with a positive slope ($\hat{\beta} = 13.39, p = 0.004$). Thus, the transparency measure that is possibly negatively correlated with $\mathcal{P}$ is positively correlated with the number of intruders.

Considered jointly, these results support the possibility that the two constituents of Mandarin two-syllable compounds are in competition: only one of them can be really transparent and pro-

ductive with respect to novel formations.

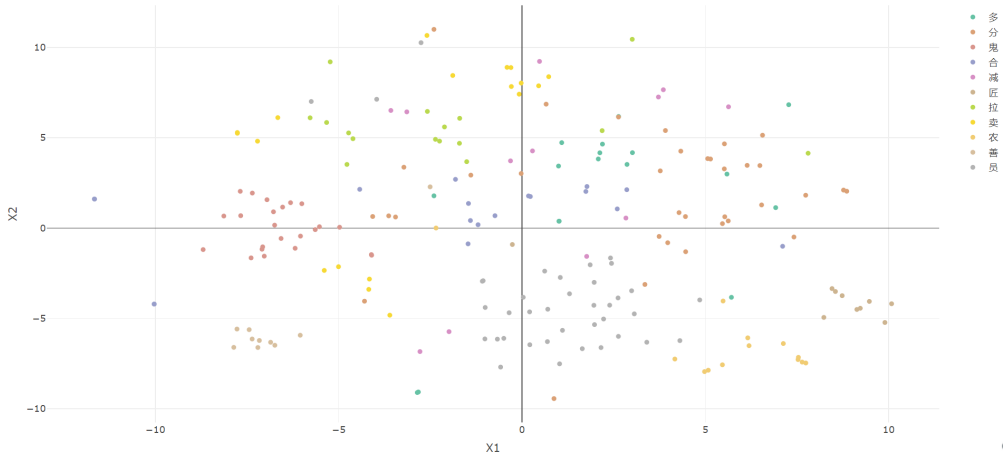

Figure 8: Compounds of the 11 intruder-free pivots with more than 10 family members in t-SNE map. These pivots show some clustering (interactive plot here).

4 The semantic geometry of suffixation

If the two constituents of compounds are in indeed in competition for productivity and semantic transparency, this raises new questions about the nature of affixes. Prefixes and suffixes are similar to pivots, but they are typically not used as independent words. It is worth noting that in English, affixes tend to give rise to affix families that are much larger than the morphological families of their base words. Inspection of the CELEX database (Baayen et al. 1995) suggests that only 3.6% of English morphemes have more than 10 family members. In other words, it is conceivable that prefixes and suffixes can be productive precisely because their base words are relatively unproductive.

What are the implications of this perspective on compounding and affixation for Mandarin suffixes? Although Mandarin is overwhelmingly a compounding language, it has a few suffixes, such as -子 (*zi4*, 'diminutive marker'), -者 (*zhe3*, 'man'), -化 (*hua4*, '-ise'), -头 (*tou2*, 'noun marker'), -们 (*men2*, 'plural marker'), -家 (*jia1*, 'certain kind of people'), and 儿 (*er2*, 'diminutive marker'). One would expect that the more Mandarin suffixes approximate suffixes in Germanic languages (see Stupak & Baayen, this volume, for German affixes), the more they should show up in clusters in t-SNE plots. On the other hand, due to the enormous productivity of Mandarin compounding, the words to which the Mandarin suffixes attach are more likely, compared to Germanic suffixes, to bring along intruders into suffixal clusters, as a consequence of the very high extent of use of Mandarin compounding. This could give rise to more diffuse clustering, and perhaps even make the suffixal pivotal systems in semantic space invisible to t-SNE.

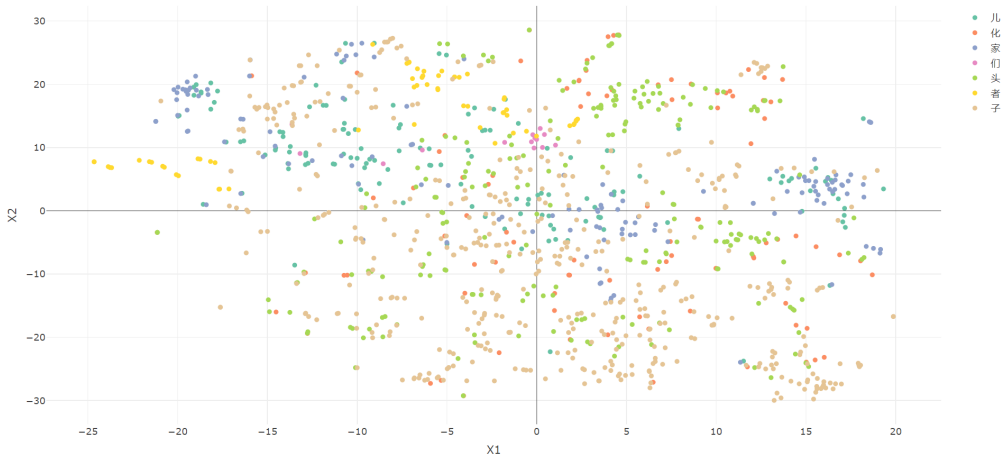

Figure 9: t-SNE map for words with the suffixes -子 (zi3, 'diminutive marker'), -者 (zhe3, 'man'), -化 (hua4,'-ise'), , 头 (tou2, 'noun marker'), -们 (*men2*, 'plural marker'), -家 (*jia1*, 'certain kind of people'), and 儿 (*er2*, 'diminutive marker'). A cluster of words with -者 is present in the upper part of the top left quadrant (in yellow). Derived words with -儿 dominate in the upper left corner along the diagonal (in dark green). -家 is represented (in blue) by three small clusters (interactive plot here).

Figure 9 presents a t-SNE map for the main seven suffixes of Mandarin. Some by-suffix clustering is visible. A cluster of words with -者 is present in the upper part of the top left quadrant (in yellow). Derived words with -儿 dominate in the upper left quadrant along the the diagonal (in dark green). -家 is represented by three small clusters along the diagonal from the top left to the bottom right (in blue). However, clusters overlap to a much larger extent than is the case for case-inflected words in Russian and Finnish (Chuang et al., this volume; Nikolaev et al., this volume), and most derivational suffixes in German (Stupak & Baayen, this volume).

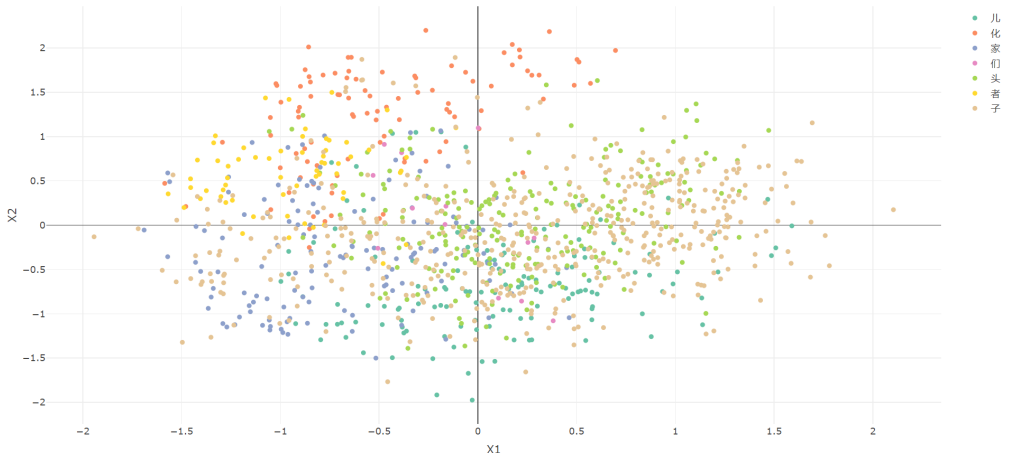

Figure 10: t-SNE 2D map of words with seven Mandarin suffixes obtained with multidimensional scaling. Words cluster more closely by suffix compared to the t-SNE clustering method, but here too clusters overlap to a considerable extent (interactive plot here).

As t-SNE, in order to best pull clusters apart, does not respect the original distances in semantic space, we also investigated the geometry of Mandarin suffixes using classical multidimensional scaling, which is designed to respect distances in the original space as much as possible in its two-dimensional projection. Figure 10 displays locally more cohesive clusters. For instance, words with -家 (*jia1*, 'certain kind of people') are located in the top left quadrant of the plane (in orange), and words with -儿 are dominant in the central bottom part (in dark green). However, also in this figure we find considerable overlap for the clusters of many pairs of suffixes. Nevertheless, we now have some evidence that Mandarin suffix families form clusters in semantic space, similar to intruder-free pivots in compounds.

The analyses of Mandarin suffixes presented thus far can be refined in two ways. First, some of the words that we included are arguably compounds instead of suffixed words. The problem is that the characters representing suffixes are also in use as independent words, albeit with somewhat different meanings. For instance, -子 functions as a suffix with diminutive semantics. However, it is also used to denote people, as in 学子 (*xue2zi3*, 'student'), and 贼子 (*zei2zi3*, 'traitor'). We therefore removed all words (355) from our suffix dataset (1093) in which the characters realize the semantics of content words, instead of their more abstract suffixal meanings.

A second refinement addresses how we carry out the clustering analysis. Thus far, we have investigated the geometry of the embeddings of the complex words themselves. But this method does not take into account the semantics of the base words. Following Shafaei-Bajestan (this volume), Chuang et al. (this volume) and Nikolaev et al. (this volume), we calculated for each suffixed word the vector that, when added to the vector of the base word, results in the vector of the suffixed word. For 儿子 (son, 'son'+'diminutive marker'), this shift vector is obtained as follows:

$$\overrightarrow{\text{儿子}} = \overrightarrow{\text{儿}} + \underbrace{(\overrightarrow{\text{儿子}} - \overrightarrow{\text{儿}})}_{\text{shift vector}}.$$

In other words, the shift vector of a suffix is obtained by subtracting the vector of a base word from the vector of the derived word.

We can now use t-SNE (or other dimension reduction methods) to investigate whether the shift vectors show clustering in the high-dimensional space of shift vectors. If a suffix contributes relatively similar shifts in meaning across the base words that it attaches to, then we may expect that the individual shift vectors for that suffix cluster in shift space.

Figure 11 presents the t-SNE map for the shift vectors of the Mandarin suffixes. Reasonably well-delineated clusters emerge from this unsupervised clustering analysis. Words with -儿 are located in the upper central part (in dark green), words with -化 dominate in the right bottom quadrant (in orange), and words with -者 cluster in the lower right quadrant along the x-axis (in yellow). This analysis provides further evidence that several Mandarin suffixes indeed contribute a specific 'shift' in meaning, and as a consequence are semantically reasonably transparent exponents of Mandarin.

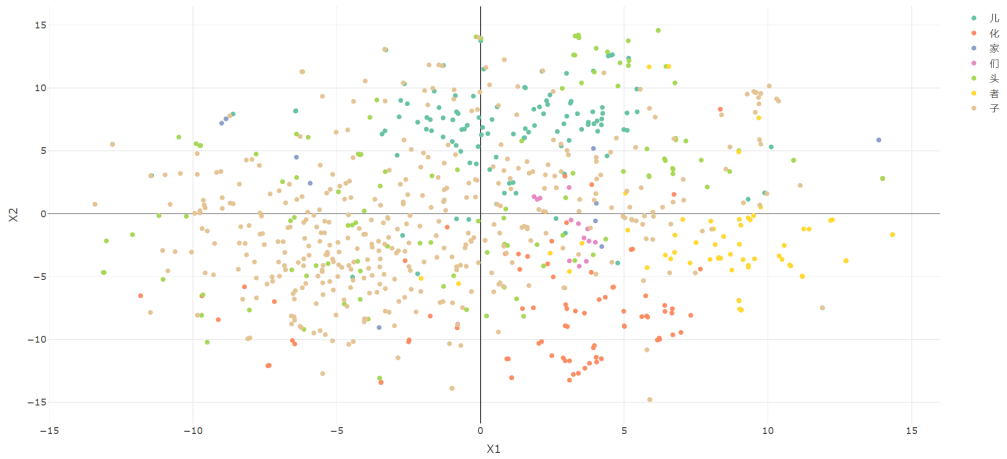

Figure 11: t-SNE map for seven Mandarin suffixes, based on shift vectors. Fairly well-delineated clusters are present. Words with -儿 are located in the upper central part (in dark green), words with -化 dominate in the right bottom quadrant (in orange), and words with -者 (in yellow) are clustered in the right bottom quadrant along the x-axis (interactive plot here).

These results raise the question of whether the suffix-specific shift vectors are possible thanks to the absence of large numbers of intruders into the suffix clusters in semantic space. To assess whether suffixes tend to be more intruder-free than compound pivots, we calculated the number of semantic intruders for pivots and suffixes based on the joint data of pivots and suffixes.[2] As suffixes and pivots have morphological families that differ substantially in size, we calculated for each pivot and each suffix the ratio of the number of intruders and the pivot or suffix family size, in order to gauge the extent to which a pivot or suffix is intruder-free. A Wilcoxon-test on the ratios indicates that suffixes may be more intruder-free than pivots ($p = 0.02$).

[2]Since pivots and suffixes may interfere with each other in a semantic space, the numbers of intruders for pivots are greater than the corresponding numbers obtained in a semantic space populated only with compounds. In general, all results presented in this study are, unavoidably, conditional on the materials that we selected for analysis.

We also calculated the semantic similarity of the pivot to its carrier word, extending the $s_{\text{piv-com}}$ and $s_{\text{nonpiv-com}}$ measures to the suffixed words, resulting in the measures $s_{\text{suf-der}}$ and $s_{\text{base-der}}$. For the suffixed words, the mean of $s_{\text{suf-der}}$ is less than $s_{\text{base-der}}$ ($t(1092) = -27.332, p < 0.0001$). However, for compounds with 31 intruder-free pivots, $s_{\text{piv-com}}$ and $s_{\text{nonpiv-com}}$, no such difference is detectable ($t(372) = -0.637, p = 0.524$).

The reduced similarity of the suffix meaning and the derived word meaning, compared to the base word meaning and the derived word meaning, makes sense. For ease of exposition, consider English suffixes such as *-ness* and *-ly*. The interpretation of words such as *smoothness* and *smoothly* depends mostly on the meanings of the base word. The semantic contribution of the suffixes is very predictable. As a consequence, the shift vectors of *-ness* and *-ly* largely preserve the semantics of their base words. We think that for Mandarin suffixes, the quantitative factors that shape the semantic lexicon have conspired to work in the same direction, albeit in a more diffuse way — an issue to which we return in more detail below.

In the light of these considerations, it is perhaps unsurprising that the semantic similarity between the suffix and the derived words ($s_{\text{suf-der}}$) appears to be smaller than that between the pivot constituent and the carrier compounds with intruder-free pivots ($s_{\text{piv-com}}$) ($t(528) = -10.621, p < 0.0001$).

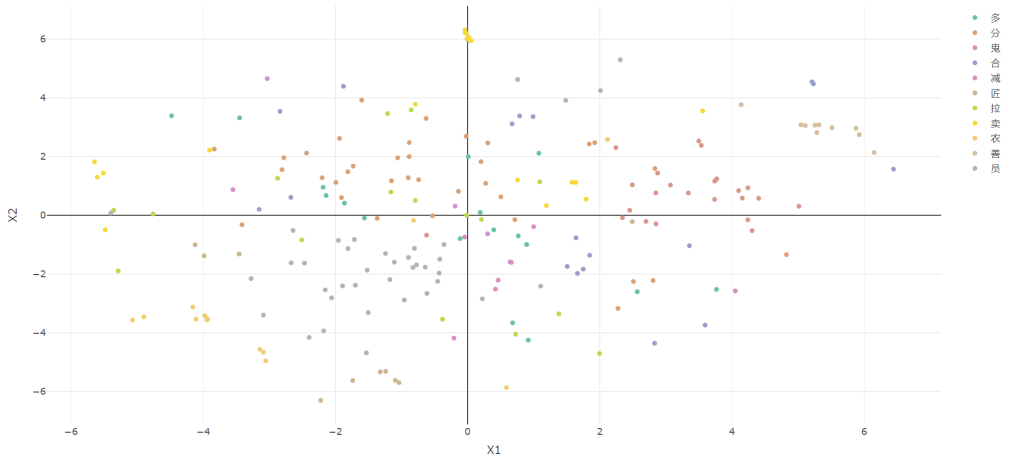

Figure 12: Compounds of the 11 intruder-free pivots with more than 10 family members in t-SNE map for their shift vectors. Of these pivots, 员, *yuan2*, 'member'; 善, *shan4*, 'virtue' ; 农, *nong2*, 'farmer'; 鬼,*gui3*, 'ghost' and possibly 分, *fen1*, 'separate' show some clustering. (interactive plot here).

The clustering of shift vectors visible for suffixed words raises the question of whether similar clustering characterizes intruder-free pivots. Figure 12 shows that clustering is present for roughly half of these pivots (员, *yuan2*, 'member'; 善, *shan4*, 'virtue' ; 农, *nong2*, 'farmer'; 鬼,*gui3*, 'ghost' and possibly 分, *fen1*, 'separate'), but not for all. This suggests that some intruder-free pivots have not undergone semantic bleaching, and do not contribute a predictable semantic change to their base words, whereas others more approximate the quantitative patterning of suffixes.

Considered jointly, our findings suggest that, at least for Mandarin Chinese, there is a gradual

transition from compounding with a large intruder ratio to compounding with few or no intruders, to affixation with very low intruder ratios. Large intruder ratios are detrimental for the productivity of pivots, in the sense that they blur out semantic regularities, and hence stand in the way of the development of pivot 'schemas' or 'constructions' that are crucial for enabling generalization to create novel complex words. Interestingly, the intruder-free pivots in our dataset are intermediate between general pivots and suffixes with respect to whether they realize similar shift vectors with respect to their base words. Some of these pivots resemble suffixes in that they do reveal some clustering of their shift vectors, but others do not show clear clustering. Their motivation with respect to their base words appears to be more on a case-by-case basis.

Finally, the clustering in t-SNE space that we observed for intruder-free pivots appears to be sufficient to break the correlation between $\mathcal{P}$ and V that governs the complete set of pivots ($p > 0.3$).

5 General discussion

This study started out with the observation that Mandarin pivots show a linear relation with negative slope for $\log V$ and $\log \mathcal{P}$, a relation that is not present for English affixes. We then provided evidence that higher estimates category-conditioned productivity are more likely for pivots that are more similar in semantic space to the non-pivots on the one hand, and to their carrier compounds on the other hand. We then followed up on a regression analysis suggesting that a high semantic similarity of non-pivot constituents to their carrier compounds is detrimental to productivity, as gauged with $\mathcal{P}$. An analysis of the geometry of compounds and suffixed words suggested that the extent to which the clusters of words sharing the same pivot are homogeneous, in the sense that these clusters are not invaded by intruders from other clusters, is a measure that predicts both semantic transparency and possibly the probability of novel forms. As a consequence, the probability that a Mandarin compound has two pivots that are both productive is quite small.

Mandarin suffixes emerged from our analyses as 'pivots' with remarkably low intruder ratios. At the other extreme, many compound pivots have high intruder ratios. In between are pivots that have relatively few intruders (which we defined as intruder-free based on a 95% confidence interval), but whose semantics are often more similar to the semantics of their carrier compounds than is the case for suffixes and their derived words, although some intruder-free pivots revealed suffix-like shift vectors. Interestingly, for the subset of intruder-free pivots, no clear correlation between V and $\mathcal{P}$ could be observed.

These observations about the geometry of semantic transparency offer a new perspective on the relation between $\log V$ and $\log \mathcal{P}$ in Mandarin. In section 2 we showed that for Mandarin pivots, this relation is characterized by a straight line with negative slope: as the potential for novel compounds with a pivot increases, the number of types that are already realized decreases. Such a negative correlation is absent for English affixation, and also for our intruder-free pivots.

The linear functional relation with negative slope for $\log \mathcal{P}(N)$ as a function of $\log V(N)$ that characterizes the majority of pivots follows straightforwardly from a simple urn model (bag of words model). Figure 13 clarifies, using the text of Lewis Carroll's 'Alice's Adventures in Wonderland', that a similar functional relation can be present also for actual structured texts (for which the urn model provides only a first approximation (see Baayen 2001, for detailed discussion)). Taking the

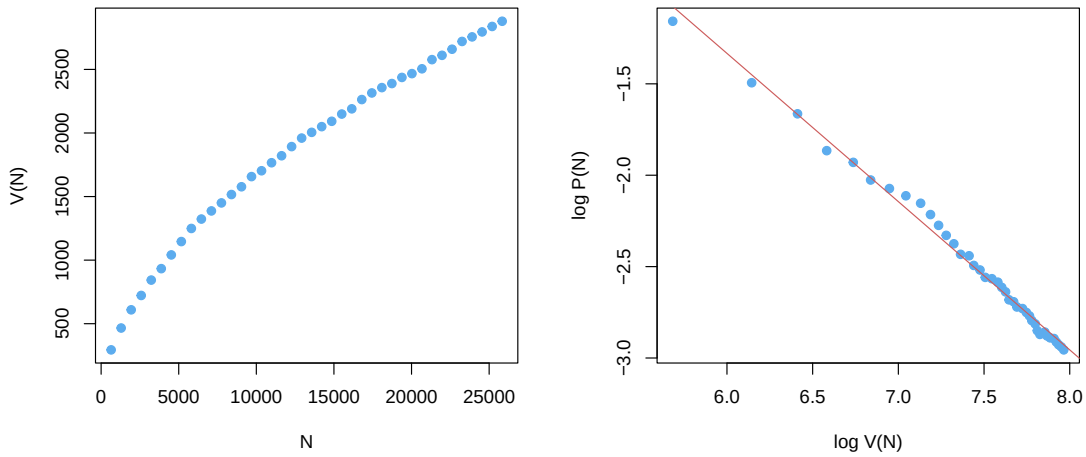

Figure 13: Dynamics of vocabulary size illustrated for Lewis Carrol's 'Alice's Adventures in Wonderland'. Left: vocabulary size $V(N)$ as a function of text size N; Right: growth rate $\mathcal{P}(N)$ is roughly a linear function of vocabulary size $V(N)$.

perspective of the urn model, pivot families resemble samples of different size $V(N)$ from the same underlying population of widely varying and heterogeneous concepts. Suffixes, by contrast, as well as some intruder-free pivots, realize more systematic changes in semantic space, changes that we gauged with the help of shift vectors. As a consequence, the morphological families of suffixes are not random samples from semantic space. This, in turn, predicts the absence of a roughly linear relation with negative slope between $\log V(N)$ and $\log \mathcal{P}(N)$.

Shen and Baayen (2021) hypothesized, on the basis of the adjectival pivots (which comprised all and only the data of their study), that the semantics of Mandarin adjective-noun compounds are more homogeneous than the semantics of affixes in English. The results of the present investigation suggest that although indeed suffixes come with distinct shift vectors, compounds sharing a given pivot tend to be more (instead of less) heterogeneous. Many pivot clusters have intruders, implying that the meaning realized with one particular pivot could also have been realized with a different pivot.

Pivots that have few intruders and that show clustering in semantic space appear to be intermediate between compounding and affixation. This suggests to us that there are two ways in which their extent of use $V(N)$ could increase over time. On the one hand, an intruder-free pivot could become more popular for creating names for concepts, at the price of semantic transparency. This development would render them more similar to pivots with large numbers of types $V(N)$, but low productivity $\mathcal{P}(N)$. On the other hand, an intruder-free pivot could undergo further semantic bleaching, reducing or even breaking the transparency of the pivot to its carrier word, and allowing the carrier word to be more transparent to its base word. This development would gradually change an intruder-free pivot into an affix.

The first kind of development can perhaps be characterized as gravitation towards 'bricolage'

(Lévi-Strauss 1962), the creative re-use of old materials for new purposes, albeit without ever completely breaking links between past and present. The positive correlation with productivity of the two transparency measures involving the pivot bear witness to the continuing links with 'the past', whereas the absence of clustered shift vectors emphasizes the re-use for new purposes.

The second kind of development, in contrast, might be profiled as gravitation towards predictability, arising thanks to shared semantic changes that are visible in our analyses as clustered shift vectors. Interestingly, predictability does not guarantee a high extent of use. For example, the Mandarin plural suffix 们 appears in some pronouns 我们 ('we'), 你们 ('you'), 他们, 她们 ('they m/f') and a few content words (e.g., 娘儿们, *niang2rmen*, 'offensive appellation for woman'; 爷儿们, *ye2rmen*, North China dialectal form for 'man'). The shift vectors for 们 cluster reasonably well, and yet the number of types in use is very small. This does not imply that 们 cannot be used to form plurals of nouns – in children's books, such plurals are occasionally used. In adult language use, however, language users are supposed to be able to infer number from the discourse context. Why pluralization enjoys a high extent of use (or profitability, in the sense of Corbin 1987) in English, but not in Mandarin, remains as an as yet unsolved "mystery of productivity" (Aronoff 1976).

Author note

This research was made possible by funding from the ERC, project WIDE-742545.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar*. MIT Press, Cambridge, Mass.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H. and Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Linguistics*, 29:801–843.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bauer, L. (1983). *English Word Formation*. CUP, Cambridge.

Bauer, L. (2001). *Morphological productivity*. Cambridge University Press, Cambridge.

Booij, G. E. (1977). *Dutch Morphology. A Study of Word Formation in Generative Grammar*. Foris, Dordrecht.

Ceccagno, A. and Basciano, B. (2007). Compound headedness in Chinese: An analysis of neologisms. *Morphology*, 17(2):207–231.

Corbin, D. (1987). *Morphologie derivationelle et structuration du lexique [Derivational morphology and lexical structure]*. Niemeyer, Tübingen.

Lévi-Strauss, C. (1962). *Savage mind*. University of Chicago.

Marle, J. v. (1985). *On the Paradigmatic Dimensions of Morphological Creativity*. Foris, Dordrecht.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen [Productivity as a morphological phenomenon]. *Forum der Letteren*, 2:110–125.

Shen, T. and Baayen, R. H. (2021). Adjective–noun compounds in mandarin: a study on productivity. *Corpus Linguistics and Linguistic Theory*.

Tarasova, E. (2013). *Some new insights into the semantics of English N+ N compounds*. PhD thesis, Open Access Victoria University of Wellington| Te Herenga Waka.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).