Comparing the semantic structures of lexicon of Mandarin and English

September 1, 2024

Abstract

This paper presents a cross-language study of lexical semantics within the framework of distributional semantics. We used a wide range of predefined semantic categories in Mandarin and English and compared the clusterings of these categories using FastText word embeddings. Three techniques of dimensionality reduction were applied to mapping 300-dimensional FastText vectors into two-dimensional planes: multidimensional scaling, principal components analysis, and t-distributed stochastic neighbor embedding. The results show that t-SNE provides the clearest clustering of semantic categories, improving markedly on PCA and MDS. In both languages, we observed similar differentiation between verbs, adjectives, and nouns as well as between concrete and abstract words. In addition, the methods applied in this study make it possible to trace subtle differences in the structure of the semantic lexicons of Mandarin and English.

Keywords: distributional semantics, mental lexicon, semantic vectors, clustering, procrustes analysis

1 Introduction

This paper presents a cross-linguistic study of lexical semantics within the framework of distributional semantics. We make use of a wide range of predefined semantic categories in Mandarin and English, extracted the FastText word embeddings of words from these categories, and compared the semantic structures of the lexicons of Mandarin and English using different techniques of dimensionality reduction, as well as graph theory and procrustes rotation.

The central hypothesis motivating the present study is that the subtle differences in how words of different semantic categories are used in Mandarin and in English is likely to be reflected in the corpus-based semantic vectors of these words, referred to as ‘embeddings’ in computational linguistics and distributional semantics (Harris, 1954; Landauer and Dumais, 1997; Shaoul and Westbury, 2010; Pennington et al., 2014; Bojanowski et al., 2017a). The underlying intuition in distributional semantics is that words that are used across similar contexts will tend to be similar in meaning. The way in which ‘similarity’ is operationalized differs across computational implementations generating embeddings. What is common to all methods is that similarity is based not only on a given word’s own specific contexts, but also on the contexts of the other words that occur in these contexts, and the contexts of these words, and so on. Thus, the semantic vectors for Mandarin and English words hold the promise to encapsulate aspects of their use that are less straightforward to detect by methods such as behavioral profiling (Divjak and Gries, 2009), even when using fine-grained annotation.

In the present study, we examine the semantics of the lexicons of Mandarin and English by means of 300-dimensional FastText vectors (Bojanowski et al., 2017a) for Mandarin and for English. Word embeddings for Mandarin and English words are relatively comparable because both sets of word embeddings were trained on corpora of Common Crawl and Wikipedia. At the lower level of the kinds of specific registers of written language sampled by these corpora, some between-language differences are expected. However, how a language makes use of (often highly language-specific) registers is a defining part of that language, especially in the light of the large volumes of data that underlie the FastText embeddings. This justifies the use of FastText embeddings to explore and compare the semantic space of Mandarin and English.[1]

The central research goals of this exploratory study are the following. The first goal is to use distributional semantics to make visible systemic similarities and differences in the semantic organization of the lexicon of Mandarin and likewise in the lexicon of English. The second goal is a methodological one: to compare unsupervised clustering methods and other multivariate statistical methods to trace similarities and differences across languages in order to obtain a better understanding of their strengths and weaknesses. The third goal is to explore the potential of Procrustes analysis within our linguistic study, aiming to compare semantic spaces across languages with a humble intent to uncover subtle cross-language similarities and differences.

The remainder of this paper is organized as follows. We first introduce the data of this exploratory study in Section 2. We explain in detail how we selected Mandarin and English words as well as the semantic categories to which we assigned these words. We make use of three different classifiers (linear discriminant analysis, support vector machines, and random forests) to validate the semantic categories, using FastText embeddings. Section 3 then reports three studies that investigate the similarities and dissimilarities between the semantic spaces of Mandarin and English. Here, we make use of three unsupervised clustering techniques:

[1] In this study, we do not consider contextualized embeddings (Bengio et al., 2000; Melamud et al., 2016; Raffel et al., 2020) for four reasons based on our experiences thus far. First, contextualized embeddings typically form clusters by word. Second, within these clusters, it is only occasionally that different senses form distinct sub-clusters. Third, it is far from trivial to understand the position of an individual contextualized embedding within a cluster given the words it co-occurs within its contexts. Fourth, since we are considering words independently of context, it is not clear to us what the advantage would be of working with a cloud of exemplars rather than with an embedding that approximates the centroid of that cloud (see also Arora et al., 2020).

multidimensional scaling (MDS), principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE). In Section 4, we examine the centroids of the semantic categories in the Mandarin and English embedding spaces, using multidimensional scaling to study distances between these centroids, and a network analysis of the cosine similarities of the centroids. Finally, in Section 5, we make use of a procrustes rotation to align the Mandarin and English embedding spaces. We then study the resulting shared semantic space using both t-SNE and MDS. The final section presents a discussion of our findings.

2 Data

This section presents the dataset that we compiled for this study. In Section 2.1, we introduce how we investigated our semantic categories and the words that we assigned to these categories, together with our selection criteria. In Section 2.2, we evaluate the quality of our semantic categories with three different classifiers, each of which is given the task to predict a word's semantic category from its FastText embedding. The statistical analysis was conducted using R version 4.2.2 (R Core Team, 2022).

2.1 Culture-specific sublexicons

We collected words for 21 partly culture-specific semantic categories for Mandarin and for English. Table 1 presents an overview of these categories and the number of words in each category. We defined these categories by hand, based on a combination of intuition, common sense, and the consultation of reference works. We proceeded as follows.

First, we consulted the *Chinese-English Bilingual Visual Dictionary* (Wilkes, 2008) and the *Modern Chinese Dictionary, 7th edition* (Dictionary Office, 2016), which are important sources for language learning. In these dictionaries, words are classified into different categories, shown in displays that bring together different exemplars such as animals, vehicles, or foods. Second, we included the prototypical members of each category by consulting frequency dictionaries (Davies and Gardner, 2013; Xiao et al., 2015), and selecting the most commonly used exemplars for inclusion in our dataset. Third, for polysemous words with senses falling into different semantic categories, we consulted Princeton English Wordnet, Chinese Open Wordnet, and the dictionaries mentioned above, and selected the dominant sense for inclusion in the data sets. Where sources diverged with respect to the dominant sense, the first author selected the sense she judged to be the most important.

We allowed for semantic categories to be populated by different numbers of words, both across categories and within categories across languages. Furthermore, we did not attempt to impose one-to-one translation equivalence for Mandarin and English words in a given semantic category. China is geographically distant from English-speaking countries, so sets of words for foods, plants, or family members, can be disjunct to a considerable extent. We have avoided including large numbers of specialized terms in our categories. There are hundreds of words for trees, but many individual language users will only have a good understanding

Semantic Categories	**Mandarin**	**English**
FOOD	205	241
PLANT	70	83
APPEARANCE artifact	77	61
HOME artifact	70	111
VEHICLE artifact	36	38
WORK artifact	26	31
BODY	102	128
ANIMAL	126	160
PERSON	393	269
SUPERNATURAL	63	61
TIME	82	47
COLOR	14	11
POSITIVE adjectives	240	266
NEGATIVE adjectives	190	203
CHANGE verbs	60	78
COGNITION verbs	71	69
MOTION verbs	82	81
PERCEPTION verbs	55	51
SOCIAL verbs	65	89
ONOMATOPOEIA	72	35
MODALS	74	37
total	2173	2150

Table 1: Number of words for the partly culture-specific semantic categories used for Mandarin and English.

of a small subset of these names. This consideration has led to focus primarily on common and generally well-known words.

Unavoidably, our list of categories and the words in these categories are far from exhaustive. However, for the purposes of the present study, the wide range of categories and the large numbers of different words taken into consideration provide a reasonable basis for investigating how in Mandarin and English, words from different categories are positioned with respect to each other in semantic space. In what follows, we document in some detail what choices we made when compiling our dataset.

We included several categories with concrete words, both animate and non-animate. The referents of these words tend to be basic entities in the natural world (as filtered through human perception and cognition) and human society. We assigned words for trees, plants, and flowers to the category of PLANT. The set of ANIMAL nouns includes both wild animals, domesticated animals, insects, fish, and also various kinds of microbes. The FOOD category comprises words for man-made foods and drinks such as *noodles*, *bread* , *soup* and *beer*, as well as words denoting different kinds of meat (e.g. *beef* and *pork* in English, 牛肉 *niú-ròu* and

猪肉 *zhū-ròu* in Mandarin). In addition, those words for animal and plant that are predominantly used to denote foods are also included in the category of food. Whether a word is predominantly used to denote food was determined by entering the word as a search term for Google Images and inspecting the images returned for that word. If the majority of images represent food rather than animals or plants in nature, the word was assigned to the category of food. For instance, we conducted searches on Google Images using the keywords *chicken*, 鸡 *jī*, and 鸡肉 *jī-ròu* on January 19, 2024, retrieving the initial 20 images for each term. Notably, we observed that out of the first 20 images associated with the term ”chicken,” only 3 depicted animals, while 19 out of 20 images for the term 鸡 *jī* and none for 鸡肉 *jī-ròu* represented animals. Consequently, we categorized the terms *chicken* and 鸡肉 *jī-ròu* as food, and 鸡 *jī* as animal.

In addition to the entities from the natural world, we considered nouns that refer to what we make and use in our social life. In the present study, we group these artifact nouns into four subgroups. The first group comprises the nouns denoting the artifacts that are used for appearance, such as clothing and cosmetics. The second group includes nouns that are related to the home. For example, both languages have words for describing the parts of a house, words for furniture (e.g. 沙发 *shā-fā* and *sofa*), and words for electronic devices used at home. The third group is comprised of the things used at work, such as computers, pens, and desks. The last group brings together vehicle nouns that denote specific means of transportation.

The category of person comprises both kinship terms (of which Mandarin has many) as well as words for occupations. Examples of kinship terms in Mandarin are 叔叔 *shū-shu* (‘father’s younger brother’) and 婶婶 *shěn-shen* (‘father’s younger brother’s wife). The words for occupations includes words such as 医生 *yī-shēng* (‘doctor’) and 教授 *jiào-shòu* (‘professor’) in Mandarin, and ‘baker’ and ‘professor’ in English.

Our dataset also includes words for human body parts. The body parts of human beings are universal, but the referents of these words may differ between Mandarin and English. For instance, in our dataset, we included 腰 *yāo*. This word does not have an exact equivalent in English. The best approximate translation in our dataset is *waist*. 腰 *yāo* usually refers to the body’s waist region, and most often describes the area where we find the lumbar vertebrae, the lower back muscles, and the corresponding tissues. In Mandarin Chinese, 腰 *yāo* is a frequently used noun compared with nouns for other body parts, whereas *waist* is less important for English users.

Words for supernatural beings can provide a window on the culture in which a language is used. Supernatural beings in Mandarin can be traced back in part to mythical tales like *Journey to the West* and *Strange Tales from a Chinese Studio*. Examples are 幽灵 *yōu-líng* (‘ghost, spirit’) and 神仙 *shén-xiān* (‘god’). But we also included 菩萨 *pú-sà*, ‘buddha’ and 观音 *guān-yīn*, ‘female buddha’, in this category. English names for supernatural beings stem mainly from monotheistic religions (‘god’, ‘angel’), but also from folklore and fairy tales (‘elves’, ‘ghosts’).

time expressions reveal how we perceive the temporal succession of days, months, and seasons. We included regular time expressions such as 分钟 *fēn-zhōng / minute*, 小时 *xiǎo-shí/ hour*, 天 *tiān / day*, 星期 *xīng-qī /week* , 月 *yuè / month*, 季节 *jì-jié / season*, and 年 *nián /year*. However, we also considered some time expressions unique to one language. In Mandarin, the year is divided into 24 parts marked by 24 special days, 节气 *jié-qì*. Every *jié-qì* has its own name featuring the season, climate, or temperature, all of which play an important role in agriculture. Furthermore, the day is divided into 12 parts, the 时辰 *shí-chén*, which traditionally regulated daily schedules. For English, time expressions include the names for the days of the week and the names of the months.

As to verbs, five subgroups were selected, ranging from verbs describing concrete actions to verbs describing abstract social and mental activities. Due to the polysemy of many verbs, we labelled the verbs according to the first sense in Chinese Open Wordnet and in English Wordnet. The set of MOTION verbs contains verbs denoting the act of moving from one place to another, such as *come*/来 *lái*, *go*/去 *qù*, and *walk*/走路 *zǒu-lù* in Mandarin. Verbs of CHANGE, such as *increase* and 增加 *zēng-jiā*, describe actions or processes that involve a change in state or condition. PERCEPTION verbs denote sensory experiences related to vision, sound, smell, taste, and touch. Typical members in this group include *see*, *hear*, and *feel* in English, and 看 *kàn*, 听 *tīng*, and 感觉 *gǎn-jué* in Mandarin Chinese. COGNITION verbs describe mental processes, thoughts, and intellectual activities such as *think* and 思考 *sī-kǎo*/认为 *rèn-wéi*. The last group of verbs contains verbs that describe SOCIAL interactions. For instance, *celebrate* and 庆祝 *qìng-zhù* are used to denote special activities at a variety of public or private events, such as parties, gatherings, and ceremonies.

We also included two sets of adjectives. Although there are many different classes of adjectives, the present study only focuses on evaluative adjectives. For both languages, we selected the adjectives with POSITIVE and NEGATIVE meanings. In English, the adjective *happy* has a positive valence, denoting a state of well-being and contentment. Likewise, the adjective 高兴 *gāo-xìng* ‘happy’ in Mandarin Chinese also has a positive connotation, denoting a comparable emotional state. We do not consider other adjectives with neutral meaning or those derived from nouns and verbs.

In addition to verbs and adjectives, we included MODAL expressions as an independent category. This category contains words that expresses the speaker’s attitude or the necessity, possibility, probability, or desirability of a situation. English modals comprise auxiliaries such as *should* and *must* and adverbs such as *certainly* and *obviously*. Examples of Mandarin modal expressions are 可能 *kě-néng* and 八成 *bā-chéng*, which translate as ‘possibly/may/might/can/could’ and ‘can/could with around 80 percent of likelihood’ respectively. 显然 *xiǎn-rán*, ‘obviously’, and 必然 *bì-rán*, ‘sure to’, are further examples of Mandarin modals.

We also included the basic color terms in Mandarin and in English. Mandarin COLOR words in this dataset are colors without the character 色 *sè*. Some words specific to Chinese culture is included in this dataset, such as 青 *qīng*, a color that falls between blue and green. Mostly, these COLOR words are used as adjectives, but they can also be used as nouns.

Furthermore, we included the most salient onomatopoeic words in both Mandarin and English. An ONOMATOPOEIA is a word that phonetically imitates, resembles, or suggests the sound it describes. ONOMATOPOEIA are used more widely in Mandarin, and have greater cultural significance. Examples of Mandarin onomatopoeia are 唧唧 *jī-jī* (‘sound of birds or insects chirping’) and 哼哧 *hēng-chī* (‘puff hard, be out of breath’); examples of ONOMATOPOEIA in English are *buzz* and *swoosh*. The Chinese and English onomatopoeia were extracted from some online word lists of onomatopoeia such as Wikipedia and Baidu

Models	Mandarin	English
Linear Discriminant Analysis	91.95%	85.90%
Support Vector Machines	91.03%	87.67%
Random Forest	77.47%	82.32%

Table 2: Classification models and prediction accuracy for held-out data.

Baike.

The lists of all Mandarin and English words used in this study are available in the supplementary materials at `Mhttps://osf.io/ge2m6/?view_only=d69d1ecb29f94566ae23342b3e25c230`. To showcase the members of a semantic category, we also list all words in the category of BODY in Appendix. These lists are not intended to be exhaustive, but rather to be sufficiently rich for our exploration of multi-category semantic profiling of the Mandarin and English lexicons.

2.2 Predictability of classes

Since the words of the categories defined in the preceding section were selected manually, we employed three supervised learning techniques, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF), to clarify whether the categories are sufficiently distinct and separable given the embeddings of the words in these categories.

We represented words' meanings with FastText word embeddings, and we evaluated class separability using cross-validation. For LDA, we made use of leave-one-out cross-validation as implemented in the `lda` function of the **MASS** package for R. For SVM (using the `svm` function from the **e1071** package, with a sigmoid kernel) and RF (using the `randomForest` function from the **randomForest** package), the original datasets were divided into an 80% training set and a 20% testing set, with the models being fitted to the training data and subsequently evaluated on the testing data. Details about the exact settings of parameters are available in the supplementary materials.

Table 2 presents the prediction accuracy for the Mandarin and English category classes using the three classifiers. Accuracies range between 77.47% and 91.95%, with the LDA model for Mandarin showing the highest success rate. The classification results demonstrate the validity of the semantic categories established in this study. Given that many words have multiple senses and we must classify them into only one category, achieving an accuracy rate of eight or nine out of ten correctly predicted instances under cross-validation lends strong support to the distinctiveness of our defined categories.

3 Relations between the semantic categories

This section investigates how the words in the different categories are distributed in the semantic spaces of Mandarin and English. We used three techniques implementing dimensionality reduction, multidimensional scaling (Borg and Groenen, 2005; Cox and Cox, 2008, MDS), principal component analysis (Pearson, 1901; Hotelling, 1933, PCA), and t-distributed stochastic neighbor embedding (Van der Maaten and Hinton, 2008, t-SNE). We performed dimensionality reduction using the **MASS** package (Venables and Ripley, 2002) for MDS with the `isoMDS()` function, the base R function `prcomp()` for PCA, and the **Rtsne** package (Van der Maaten and Hinton, 2008) for t-SNE with the `Rtsne()` function. Our aim is twofold: first, to verify that words cluster by semantic category, and second, to gain insight into how semantic categories are positioned with respect to each other.

Multidimensional scaling is a classical technique that seeks to stay faithful to the Euclidean distances when projecting the 300-dimensional FastText vectors into a low-dimensional space. This method is extensively used in quantitative linguistic research, such as typological studies and construction grammar (Black, 1973; Gandour and Harshman, 1978; Fox et al., 1995; Croft and Poole, 2008; Levshina, 2015, 2016; der Klis and Tellings, 2022).

Principal components analysis places observations in a new coordinate system, such that the first axis (principal component, henceforth PC) explains the largest part of the variance in the data, and the last axis the least variance. PCA is also widely applied in linguistic studies (Baayen et al., 1996; Laakso and Smith, 2007; White et al., 2018; Musil, 2019).

Against the background of the results obtained with MDS and PCA, we then proceed to use t-SNE. T-SNE is a relatively novel method (Van der Maaten and Hinton, 2008) that can be seen as a nonlinear version of multidimensional scaling. It relaxes the assumption that distances in the original high-dimensional space should be reflected as faithfully as possible in the low-dimensional projection of this space. T-SNE is described as optimal for finding and visualizing clusters, if clusters are actually present in the high-dimensional space. This unsupervised method has been applied to linguistic research recently. Asgari and Schütze (2017) presents t-SNE visualizations of past tenses in five languages. der Klis and Tellings (2022) compares MDS with LLE (local linear embedding) and t-SNE, claiming that MDS can ‘capture the main sources of cross-linguistic variation’, whereas LLE and t-SNE are better at finding clusters. In addition, Shen and Baayen (2023) and Stupak and Baayen (2023) made use of t-SNE to study the semantics of inflection, derivation, and compounding respectively.

There are other unsupervised clustering algorithms that have been put forward as alternatives to t-SNE. In Appendix 2, we illustrate that the clusterings for Mandarin produced by UMAP and PaCMAP are not superior to, or more insightful than, the clusterings produced by t-SNE. As the focus of our study is not on evaluating different unsupervised clustering methods, we do not provide further discussion of these methods.

In the following subsections 3.1 and 3.2, we first present the word clusterings in the Mandarin semantic spaces and then investigate the English semantic spaces.

3.1 Exploring Mandarin semantic space

An analysis using multidimensional scaling revealed some clusterings by semantic category on the second and third dimensions of the reduced 3D space, and Dimensions 2 and 3 are shown in Figure 1.[2] Dimension 2 contrasts nouns (mostly on the left) with words from other parts of speech (on the right). Nouns referring to PERSON are located in the upper center. The different syntactic behaviors of nouns, verbs, adjectives, and adverbs are clearly picked up by the FastText word embeddings. Dimension 2 also appears to be reflecting differences in Arousal ($r = -0.40, t(1007) = -13.92, p < 0.0001$), as gauged by the arousal norms of Xu et al. (2022).

Words for ANIMAL, FOOD, PLANT, COLOR, and TIME are located mainly in the lower left quadrant of Figure 1. Words for artifacts (HOME and APPEARANCE) are found predominantly in the upper left quadrant. The relative locations of these clusters reveal that semantically similar categories are positioned close to each other in semantic space. Most of the POSITIVE and NEGATIVE adjectives are in the lower right quadrant, and most verbs are found in the upper right quadrant. The ONOMATOPOEIA (light blue) form an independent cluster in the lower left of the lower right quadrant.

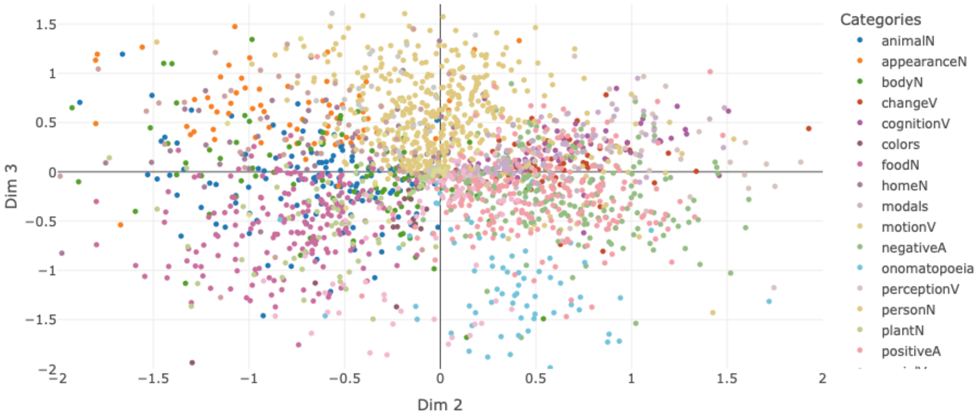

Figure 1: Clustering of Mandarin words belonging to 21 categories with MDS dimensions 2 and 3. For an interactive plot, please click here.

The first dimension reflects the difference between words with simplified characters or without simplified characters. This differentiation is largely independent of the semantic categories, performs hardly better than a baseline classifier that always predicts the majority class (accuracies: 21.6% and 18.1%).

Figure 2 presents a scatterplot for the first and third principal components of a PCA orthogonalization of the semantic space. PC1 distinguishes nouns (on the left) with words from other parts of speech (on the right). The concrete nouns referring to entities in nature are positioned on the far left. Words for TIME, artifacts (APPEARANCE, HOME, WORK, VEHICLE), and SUPERNATURAL beings are mostly on the left side but

[2] We will interpret MDS dimension 1 and PCA dimension 2 later, as they are particularly noteworthy.

close to the y-axis. The group of PERSON nouns is located in the central area of this figure. Among the groups of nouns, the category of PERSON nouns is closest to the verbs and adjectives on the right side. PC1 is also correlated with arousal ($r = 0.42, t(1007) = 14.75, p < 0.0001$), again using the ratings of Xu et al. (2022). PC3 is somewhat correlated with the valence ratings provided by Xu et al. (2022) ($r = -0.12, t(1007) = -3.86, p = 0.0001$). More POSITIVE words have higher values on these dimensions. However, this effect is largely restricted to the words for PERSON.

MDS and PCA both show a similar split into two groups, across all semantic classes, on one of their dimensions (MDS Dim1, PCA Dim2). What motivates this split is that FastText word embeddings for Mandarin were trained on corpora with both simplified and traditional Chinese characters even though this study is based on simplified Chinese. Simplified Chinese 简体中文 is the official orthography of Mainland China. To reduce illiteracy, the government of People's Republic of China simplified 2274 characters from the 1960s to the 1980s. Meanwhile, traditional Chinese 繁體中文 is still used in a lot of regions outside Mainland China, such as Taiwan, Hong Kong, Macao, Singapore and Malaysia. Chinese speakers all over the world can communicate with each other online (on platforms such as YouTube, TikTok, and Xiaohongshu) with these different writing systems. This communication is facilitated by the fact that there is a subset of characters that is used in both simplified Chinese and traditional Chinese: the characters that have never been simplified. FastText embeddings reflect co-occurrence similarities between words. Embeddings of words often collocating together will be driven closer (Rong, 2014; Bojanowski et al., 2017b). When the same words are written with different characters, their collocational profiled are affected, resulting in different FastText vectors. The simplified characters predominantly come from texts written mainly in simplified Chinese, and hence they are more likely to have other words also written with simplified characters and unchanged characters as collocates. Those characters that have never been simplified and which can be used in both simplified and traditional Chinese, have different collocational profiles, as they co-occur not only with other unchanged characters, but also with both the simplified characters and their traditional Chinese counterparts. As a consequence, the unchanged characters occur in orthographically more diversified texts and have more complex neighborhood profiles. The different usage patterns of unchanged characters and simplified characters are reflected in the FastText embeddings and are prominantly visible in the MDS dimension 1 and PCA dimension 2.

Compared to MDS (See Figure 1) and PCA (See Figure 2), t-SNE finds better semantic clusters, as can be seen in Figure 3 [3]. Nouns predominantly cluster on the left, whereas words for other parts of speech are predominantly situated on the right, which is consistent with the MDS and PCA clusterings. However, particularly the classes of nouns are very well separated. Words for entities in nature are situated in the lower left, words denoting artifacts created by human beings are in the upper left, and nouns for persons are located in the upper central area. On the upper right, groups representing verbs and adjectives show considerable overlap. On the lower right, we find an isolated cluster of ONOMATOPOEIA.

[3] UMAP is a popular alternative to t-SNE but that t-SNE, for our data, succeeds in better separating the clusters.

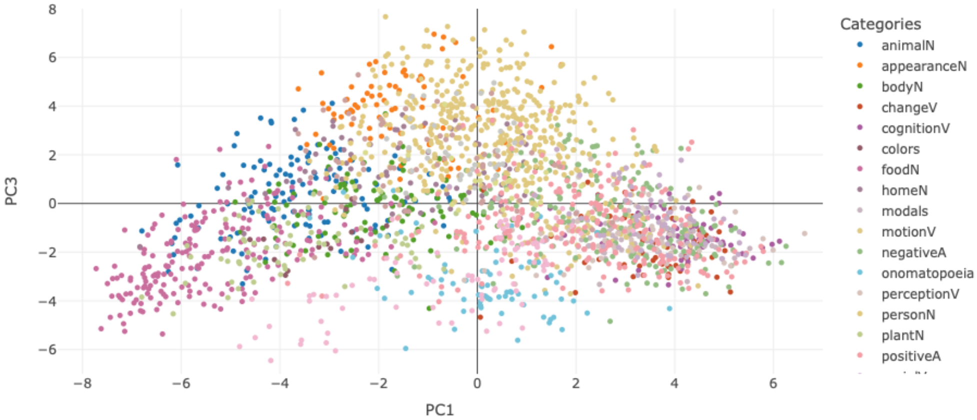

Figure 2: Clustering of Mandarin words belonging to 21 categories using PCA. For an interactive plot, please click here.

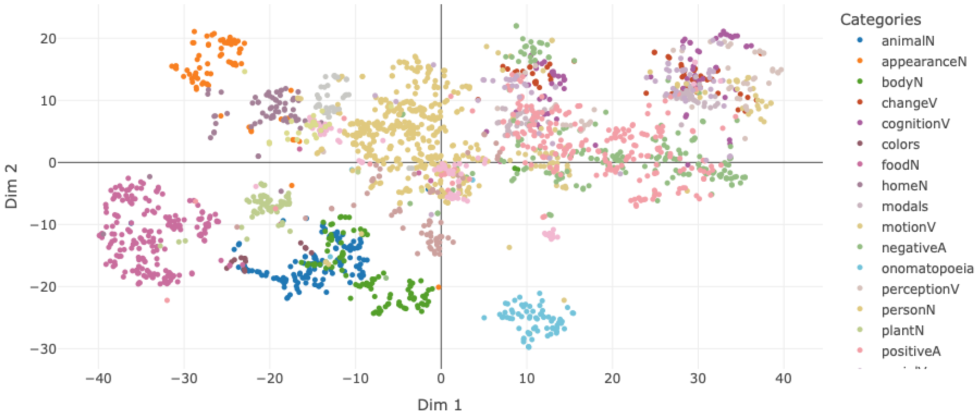

Figure 3: Clustering of Mandarin words belonging to 21 categories using t-SNE. For an interactive plot, please click here.

Although in general semantic classes are well separated, there are some areas where there is considerable overlap. The first overlapping area is positioned around the coordinates (-14, -17), where words for animal and words for body parts co-occur. Time expressions are found in three regions. One cluster is located around (-14, 6), with overlap with nouns for work. A second cluster is present to the lower right of the origin. Here, we find time expressions such as 节气 *jié-qì* ‘24 special days in a year featuring the season, climate, or temperature’, as well as words for days of the week and names of months. The third group comprises expressions such as 时辰 *shí-chén*, a traditional unit of time equal to two hours.

Second, adjectives conveying evaluative meanings do not separate into two clusters, suggesting that positive adjectives and negative adjectives in Mandarin Chinese have similar semantics. We shall see below for English that positive and negative adjectives are well separated, indicating that the result for Mandarin adjectives is not necessarily an artifact of using embeddings.

Third, the five categories of verbs do not form separate clusters, suggesting that the verbs in our dataset exhibit a greater degree of polysemy than the concrete nouns. This pattern of result is consistent with the results of the LDA analysis. The noun classes have the highest prediction accuracy (with an average accuracy of 95.36%), whereas the prediction accuracy of the verb classes is 5 to 15% lower than the overall prediction accuracy (with an average accuracy of 78.67%).

3.2 Exploring English semantic space

To explore the similarities and dissimilarities of the English semantic categories, we followed the same analytical steps as for Mandarin. In what follows, we present the cluster analyses with MDS, PCA, and t-SNE, respectively.

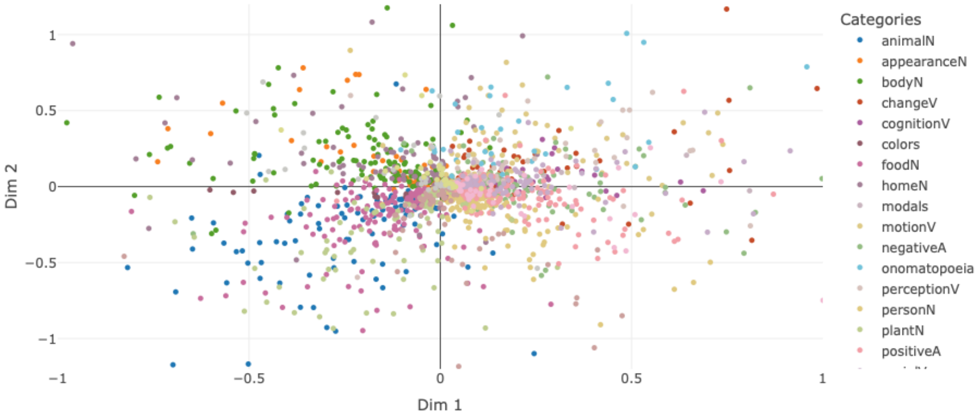

Figure 4: Clustering of English words belonging to 21 categories using MDS. For an interactive plot, please click here.

Figure 4 presents the English words in the plane spanned by the first and second dimensions of a 3D MDS analysis. Some differentiation between semantic categories is visible, but at the same time, there is substantial overlap. Nouns are located more to the left, and verbs and adjectives are found more to the right. The first dimension is somewhat correlated with arousal ($r = -0.14, t(1838) = -6.1932, p < 0.0001$), using the norms of Warriner et al. (2013). The second dimension shows a modest correlation with valence ($r = 0.09, t(1838) = 3.9596, p < 0.0001$). The third dimension (not shown) is somewhat correlated with arousal ($r = -0.14, t(1838) = -6.0755, p < 0.0001$), valence ($r = 0.08, t(1838) = 3.515, p = 0.0005$) and dominance ($r = 0.13, t(1838) = 5.7803, p < 0.0001$).

Figure 5 shows a scatterplot of the first and second principal components. On the left side, the nouns predominate but the nouns for person are mostly located in the lower right quadrant. Verbs and adjectives are positioned above the person nouns. PC1 is somewhat correlated with the ratings of arousal ($r = 0.18, t(1838) = 7.8076, p < 0.0001$), valence ($r = -0.07, t(1838) = -2.9329, p = 0.0034$) and dominance ($r = 0.05, t(1838) = -2.5325, p = 0.01$) provided by Warriner et al. (2013). PC2 is correlated with arousal ($r = -0.07, t(1838) = -2.992, p = 0.0028$) and valence ($r = -0.10, t(1838) = -4.3681, p < 0.0001$). PC3 (not shown) is weakly correlated with dominance ($r = -0.11, t(1838) = -4.8798, p < 0.0001$), valence($r = -0.08, t(1838) = -3.345, p = 0.0008$), and arousal ($r = 0.07, t(1838) = 2.8995, p = 0.0038$). This survey of correlations clarifies that none of the first three principal components reflects one particular dimension of emotionality.

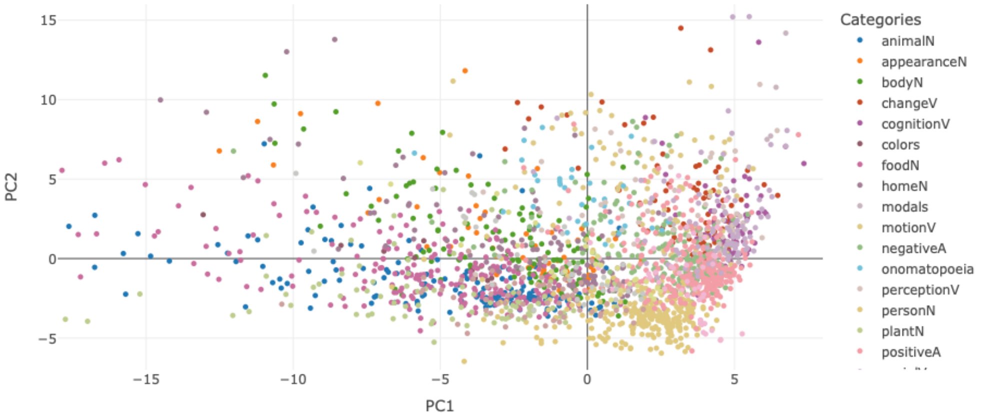

Figure 5: Clustering of English words belonging to 21 categories using PCA. For an interactive plot, please click here.

Figure 6 presents the t-SNE clusterings of the English semantic categories. Compared to the MDS and PCA visualization, the t-SNE algorithm separates the noun categories into clearly different groups. In the lower left quadrant, words for supernatural beings (near (-19, -8)) and person (around (-7, 22)) form two clusters. The group of color words is near the x-axis on the left side (-35, 5), which is close to the groups of words for plant (-31, 11) and animal (-22, 18) in the upper left quadrant. The negative adjectives *dark* and *pale* cluster with the color words. The plant name that is closest to the color words is *violet*, unsurprisingly, as this word is also used as a color word. The fact that words such as *violet* and *orange* denote both colors and plants may help explain why the color words are positioned most closely to the plant words.

Also on the left, the words for animal (dark blue) are near the x-axis (-20, 0). The animal words closest to the cluster of food nouns (-20, 20) are those referring to seafood. Nouns for animal that are not raised for food are situated on the lower part of the cluster. Below the cluster of animals, we find a cluster of supernatural beings (-20, 9). Nouns for person form an independent cluster near (-10, 20). The four groups of artifacts (nouns related to appearance, home, work, vehicle) cluster around (-2, 8). The nouns for body parts (green) are positioned highest along the y-axis (0, 28). English evaluative adjectives are well separated: positive adjectives are positioned higher in the second dimension, whereas negative adjectives are situated lower in the second dimension. The onomatopoeia form a small elongated cluster mostly above the positive adjectives. The five categories of verbs are clustered at the right-hand side of the plot (25, 0), but do not show clear between-category clustering. Modal verbs and adverbs are also found in this cluster.

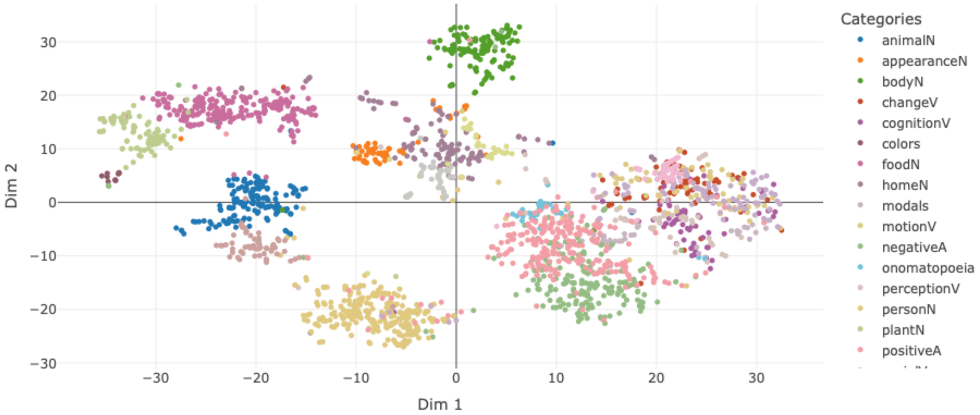

Figure 6: Clustering of English words belonging to 21 categories using t-SNE. For an interactive plot, please click here.

4 Comparison of Mandarin and English: relative positions of the clusters

This section compares the semantic clustering of Chinese and English words using different methods. The first subsection zooms in on the coordinates of the centroids of the semantic categories that were presented in Section 3. By abstracting away from the considerable overlap between categories within languages, we can bring to the fore what is similar and different between Mandarin and English. In the second subsection, we first calculate the centroid vectors of all semantic categories, and then use MDS and cosine similarity to visualize the distance and network between semantic categories.

4.1 Comparing the centroids of the semantic vectors of MDS, PCA, and t-SNE

Figures 1-6 provide full details on all words, but this makes it difficult to come to grips with the relative positions of the different categories. In what follows, we therefore focus on the centroids of the semantic categories and compares their relative positions in Mandarin and English, calculated from words' coordinates in the abovementioned figures. This considerably facilitates comparisons between languages and methods.

Figure 7 presents MDS, PCA, and t-SNE plots for Mandarin (left panels) and for English (right panels). The colors in these plots represent the parts of speech of each category. We colored nouns (that end with "N") as orange, verbs (that end with "V") as purple, and adjectives (that end with"A") as blue. As to those without uniform part of speech in both languages (that end with"O", "Others"), these are presented in green.

All sub-figures of Figure 7 distinguish nouns (in the left hand side of the scatterplots) from other categories (which are found more to the right). Within the group of nouns , the categories for ANIMALS and BODY parts are close to each other for Mandarin (left panels), but not for English (right panels). By contrast, the nouns for PERSON are relatively isolated in English (right panels). The verb categories (MOTION, CHANGE, SOCIAL, COGNITION, PERCEPTION), MODAL, and adjective categories (POSITIVE and NEGATIVE) cluster together for both languages, perhaps more tightly so in Mandarin. For English, ONOMATOPOEIA are also positioned near these categories, whereas for Mandarin, ONOMATOPOEIA are positioned at a substantially greater distance. TIME expressions appear at the right hand side of the English plots, but in Mandarin, they appear in the center, further away in the horizontal dimension from the verbs and adjectives.

Within the five sub-categories of verbs, MOTION verbs sometimes slightly move out of this cluster, as can be seen in the t-SNE plot for Mandarin and the MDS and PCA plots for English. In Mandarin Chinese, POSITIVE adjectives and NEGATIVE adjectives are very similarly positioned, especially on the vertical axis. For English, POSITIVE and NEGATIVE adjectives separate somewhat more along the vertical axis.

In both languages, COLOR words are positioned close to the words for ANIMAL, PLANT, and FOOD. In Mandarin, but not in English, the BODY centroid is also positioned fairly close to the COLOR centroid.

4.2 Comparing category centroids obtained with averaging

Thus far, we have used PCA, MDS and t-SNE to present words and category centroids in a three-dimensional space. In this section, we complement these statistical methods with an inspection of the average vectors for each of the 21 categories. The dataset that is obtained in this way contains 21 300-dimensional vectors for Mandarin, and another 21 300-dimensional vectors for English.

The average semantic vectors were obtained by summing the vectors of all words within a specific category and dividing the sum by the total number of members within that category. These average semantic vectors represent the collective semantic features of each category.

We analyzed the dataset with by-category mean vectors in two ways. In order to come to grips with

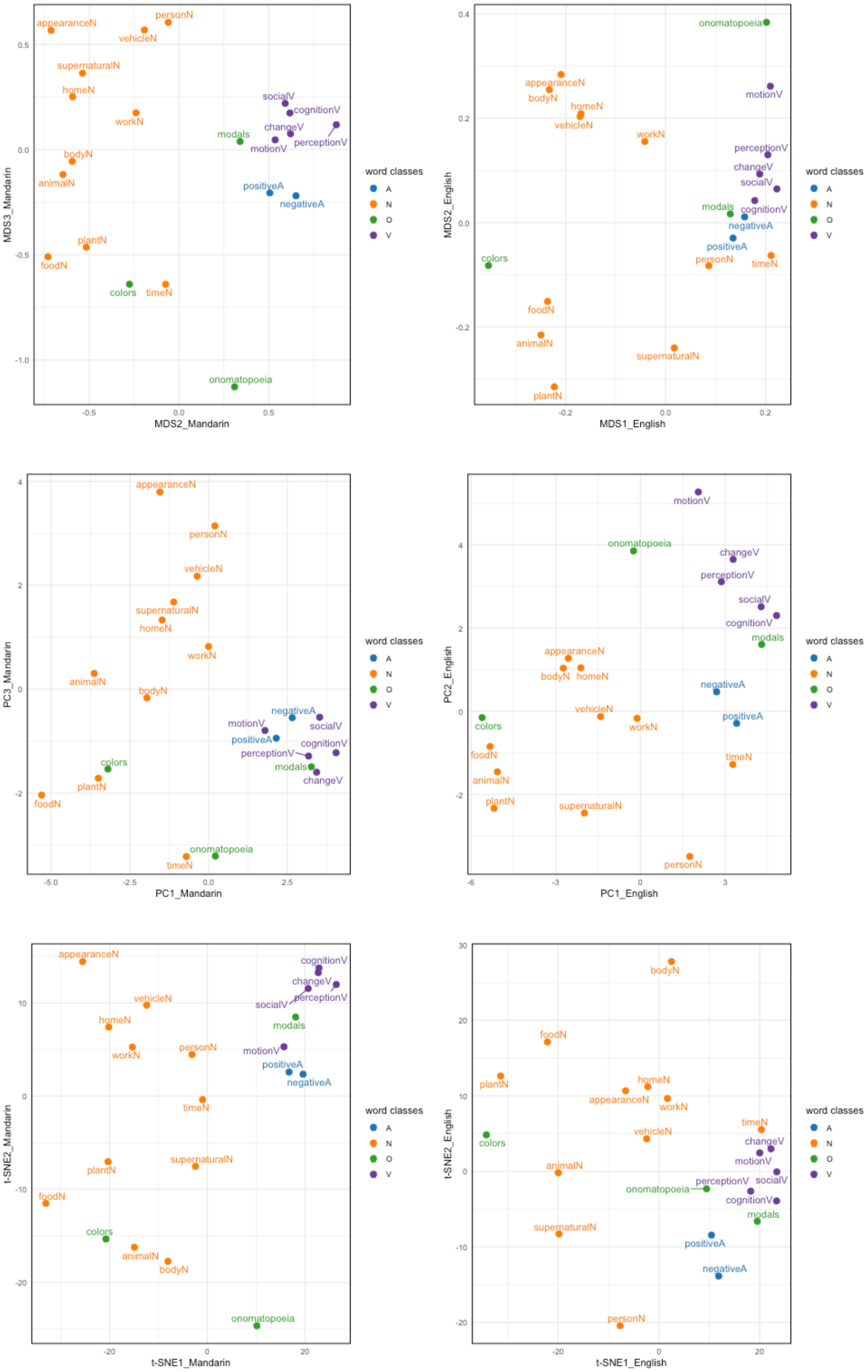

Figure 7: Scatterplots for the MDS, PCA, and t-SNE category centroids in Figures 1, 2, 3, 4, 5, and 6 for Mandarin (left panels) and for English (right panels). MDS1 and PC2 are not shown as these dimensions are captured by the aspect of traditional/simplified Chinese, as explained in Section 3.1.

distances between the centroids, we used multi-dimensional scaling. In order to trace how similar centroids are with respect to their orientation, we calculated the cosine similarities for all pair of centroid vectors, and used methods from network science (graph theory) for visualization.

Figure 8 presents the MDS plots of how by-category average semantic vectors are positioned in Mandarin (upper left panel) and in English (upper right panel). Several shared features can be observed in the left and right panels of Figure 8. First, nouns are well-separated from verbs and adjectives. Second, the categories of COLOR words and ONOMATOPOEIA are outliers in both languages. Furthermore, in Mandarin, TIME expressions are relatively independent as well. Third, in both languages, MODAL expressions are close to both verbs and adjectives, and adjectives are closer to verbs than to nouns.

The scatterplots also point to differences between the two languages. In Mandarin Chinese, the verb categories cluster less densely compared to English. This suggests that the verbs in English are characterized by a larger degree of polysemy. One possible explanation is that although single-syllable words in Mandarin Chinese are highly polysemous, multi-syllable words, which constitute 80.7% of our verbs, have far fewer senses. A complementary consideration is that English verbs, many of which in our dataset are monomorphic, are confounded with particle verbs. As a consequence, the embeddings of English verbs provide a blend of many different senses, which renders differentiation between different semantic categories less precise.

Another difference between Mandarin and English concerns the closest neighbor categories of the COLOR words. In Mandarin, the closest categories are those with words for PLANT, ANIMAL, AND SUPERNATURAL beings. In English, the words for HOME and APPEARANCE nouns are closest neighbors.

In order to assess similarities in orientation of the centroid vectors, we calculated all pairwise cosine similarities of the average vectors, resulting in two 21*21 matrices of cosine similarities. We transformed these real-valued matrices into adjacency matrices, with categories labeling rows and columns, and with an edge between two categories whenever their cosine similarity exceeded the 7th decile of the distribution of cosine similarities. Using the **igraph** package for visualization, we obtained the graphs shown in the lower half of Figure 8. (For a network analysis in which embedding-based similarities are used as connection weights, see Chen (2022).)

In both graphs, noun categories form one large cluster, and the verb categories another large cluster. The adjectives cluster with the verbs for both Mandarin and English, and for both languages, TIME expressions are completely unconnected.

The graphs also bring to light some interesting differences between the two languages. First, the ONOMATOPOEIA are integrated with the verbal cluster in English, linking up to motion and perception verbs, whereas they form a singleton cluster for Mandarin. In Mandarin, words for ONOMATOPOEIA behave like an adverbial. For example, 砰地一声 in Example (1) is equivalent to *with a loud bang* in English. In most cases, Mandarin ONOMATOPOEIA are not used as verbs. In English, some verbs encode both an action and the associated sound, as illustrated in Example (2).

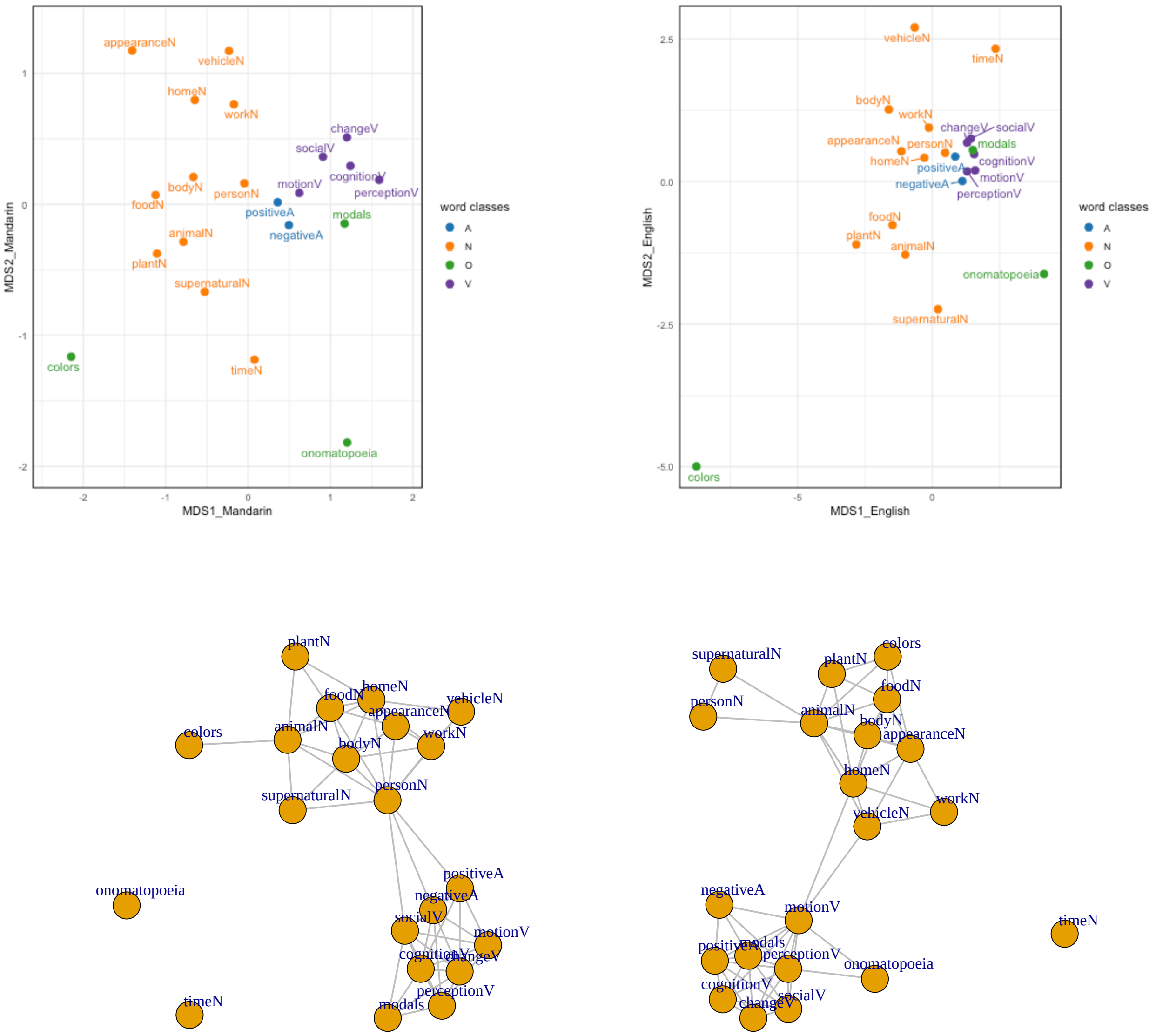

Figure 8: Analyses of category centroids for Mandarin (left) and English (right). Upper panels: scatterplots for distance, based on MDS. Lower panels: networks for angle. Vertices are connected when the cosine similarity exceeds the 7th decile of the distribution of cosine similarity.

(1) 砰地一声，一块陨石坠落在地。
pēng de yī shēng ， yī kuài yǔn-shí zhuì-luò zài dì 。
‘With a bang (loud sound), a meteorite fell to the ground.’

(2) She banged her fist angrily on the table.

Second, person nouns are linked up in very different ways in the two languages. In Mandarin, person is the pivotal category linking nouns with adjectives and social verbs, suggesting that social interaction and adjectival evaluation are important for this language. In English, by contrast, the category of person nouns has a marginal position in the cluster of nouns, with links only to supernatural beings and animal. This suggests that in English, agency or animacy is an important component of person nouns.

Third, motion verbs in English provide the only links from the verb cluster to the noun cluster, connecting up to vehicle and home nouns. In Mandarin, the motion verbs are shielded from the noun cluster by the social verbs and the adjectives (positive and negative). This suggests a stronger link in English for motion and means of transportation.

Fourth, Mandarin color terms only have high cosine similarity with animal words, but color terms in English show high similarity with a wider range of categories: nouns for plant, animal, food, and appearance .

5 Procrustes analysis of centroid vectors

The exploratory analyses presented thus far point to many similarities and some differences in the constellations of semantic categories in the distributional spaces of Mandarin and English. As a final step, we make use of a procrustes analysis to clarify whether the two distributional spaces can be mapped onto each other relatively well.

The idea of a procrustes analysis is that if two configurations of points are very similar, than if one makes sure they have the same size, and that they have the same orientation, then a rotation should suffice to line up the points of one space with those of the other. For instance, consider two leaves of the same oak tree, which are very similar in shape, but might differ in size. As a first step, we scale so that sizes are now identical. Next, we ensure that the centers of the leaves are aligned, and that they are properly oriented in the same way. Finally, we rotate one leaf so that it is on top of the other. For two oak leaves, only marginal differences should remain. By contrast, when comparing an oak leaf with a maple leaf, the procrustes rotation will not be very precise.

The sum of the squared residuals between the observed and predicted locations of pairs of points is used as metric of association. Its significance is assessed with a permutation procedure (see, e.g., Peres-Neto and Jackson, 2001) that scrambles the order of the rows in one of the matrices. If the pairs of points have the same geometrical shapes, than the association metric (a kind of correlation) should be high, and far out in the right tail of the distribution of metrics for randomized data.

A procrustes analysis requires that at least a good number of points in the one space are paired with the corresponding points in the other space. For our words, setting up such paired observations is not feasible, due to words in one language having multiple translation equivalents in the other. We therefore consider the category centroids, which are paired by language, and subject these to a procrustes analysis.

We carried out our analyses using the `protest` function from the **vegan** package (Oksanen et al., 2022). We used a symmetrical procrustes analysis, rather than an asymmetrical one, as we have no reason to give preference to one language over the other. Two analyses were carried out, one for distances and one for cosine similarities. As it is advisable to have more observations than dimensions, we reduced the number of dimensions to 10, using multidimensional scaling for the distances, and principal component analysis for the cosine similarities. The correlation metrics for both analyses were high (0.88 and 0.94), and always higher than the corresponding metrics calculated for 1000 analyses with randomly permuted data. This result dovetails well with the high degree of similarity between Mandarin and English observed in the preceding sections.

Figure 9 presents the residuals for the two analyses, which are informative about which categories are most difficult to align.

A comparison with the scatterplots in the upper half of Figure 8 is useful for understanding the stress in the procrustes mapping based on distances. Here, we focus on the four largest residuals. In English, the distance between perception verbs and negative adjectives is very small, in Mandarin these two categories are further apart. Furthermore, in Mandarin, appearance nouns and perception verbs are far apart, whereas in English, they are close together. In English, the work nouns are positioned close to the verbs, but in Mandarin, they are somewhat further away from the cluster of verbs. When evaluated in terms of distances, it is these four categories that emerge as being the most language specific.

A comparison with the graphs in the lower half of Figure 8 is helpful for understanding the large procrustes residuals based on cosine similarities. The time category has the highest residual. Interestingly, in both the Mandarin and English networks, time is an outlier. The procrustes analysis, however, suggests that the time category is an outlier in rather different ways in the two languages.

The next largest residual is for color words. In the graph for Mandarin, color links up only to animal, whereas in English, it links up to four categories (plant, animal, food, appearance). This difference is picked up by the procrustes analysis.

The large residual for person nouns is perhaps unsurprising, given that in the Mandarin graph, person is a hub linking the nominal and verbal clusters, but in English, this category is more peripheral.

The perception verbs also have a large residual in this analysis. In English, but not in Mandarin, perception verbs link up to onomatopoeia. The procrustes analysis does not report severe stress for the onomatopoeia, but clearly cannot map both categories jointly with high precision.

In order to compare the locations of Mandarin and English words in the same semantic space, we carried out an asymmetic procrustes analysis and used the `predict ()` function to rotate the Mandarin words into the English space. The result is summarized and visualized in Figure 10.

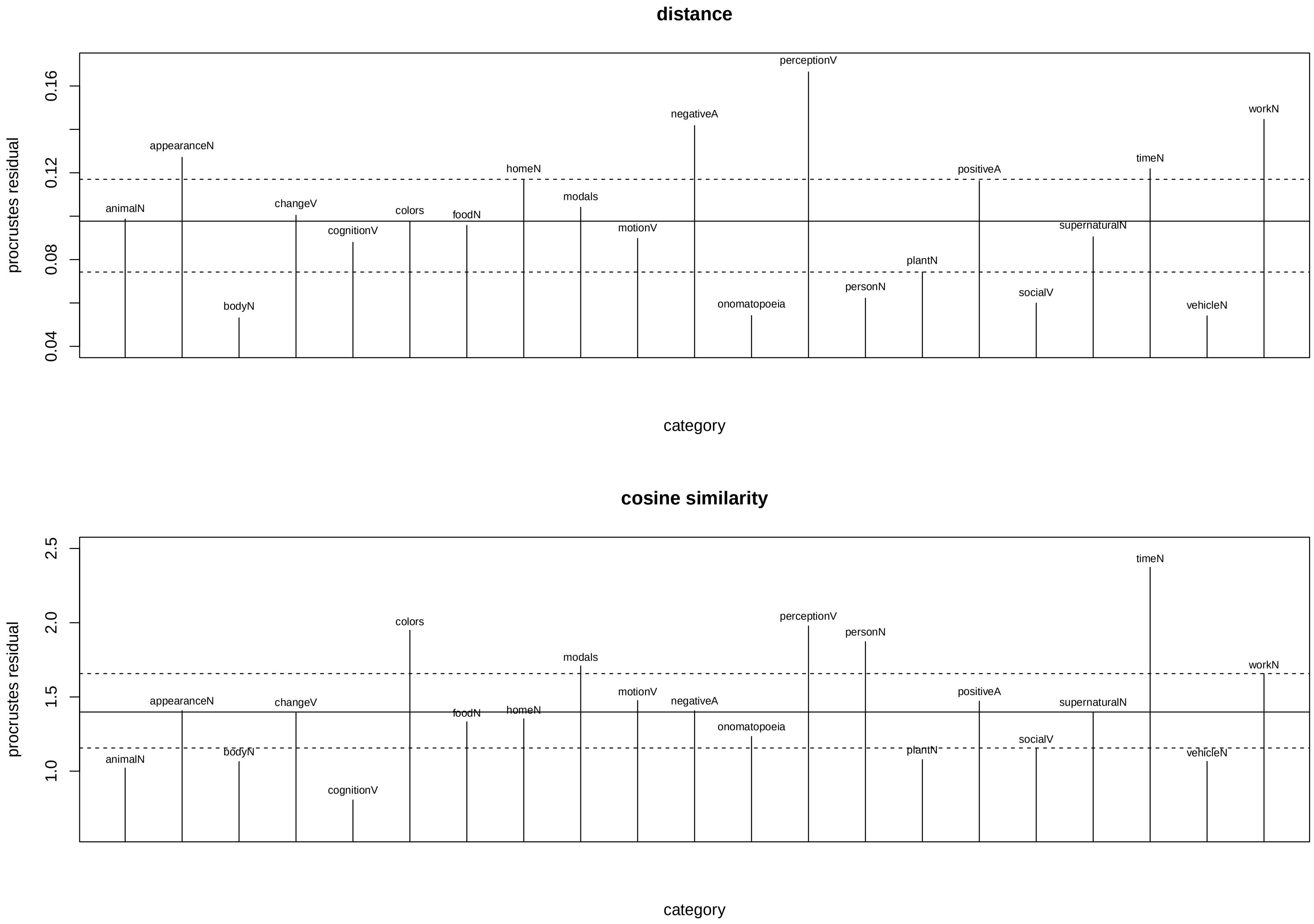

Figure 9: By-category residuals of procrustes analyses of Mandarin and English, using distances (top panel) and cosine similarities (bottom panel). The dashed lines denote the first and third quartiles.

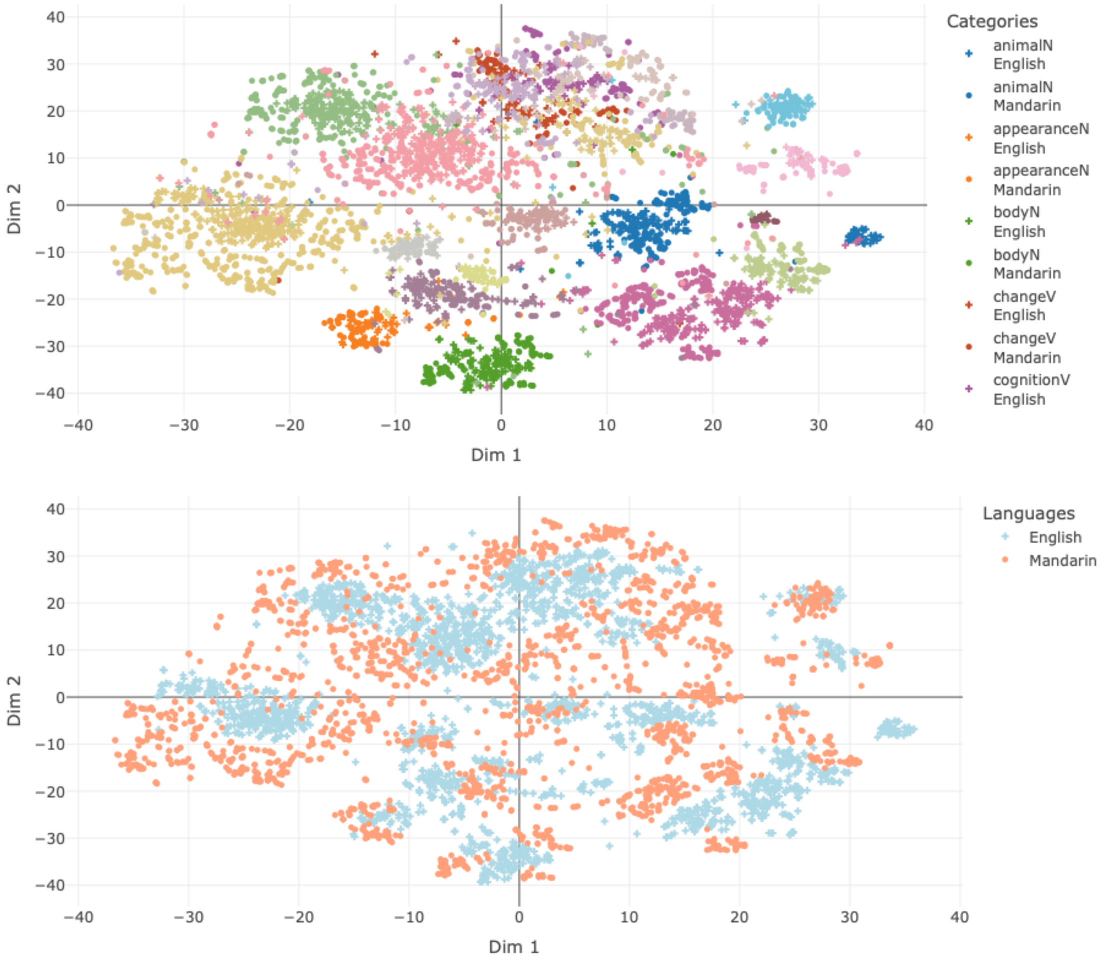

Figure 10: T-SNE scatterplots of original English and rotated Mandarin semantic vectors. Upper panel: color coding by semantic category, lower panel: color coding by language.

The upper panel of Figure 10 plots the Mandarin words (represented by dots) and the English words (represented by $+$) using color coding to highlight semantic categories (for an interactive plot, please click here). Most of the semantic categories are well differentiated in the two-dimensional t-SNE plane. An LDA analysis with leave-one-out cross validation reached an accuracy of 90.33% for predicting semantic category from the embeddings in the shared space. This indicates that the procrustes analysis is of high quality.

The lower panel of Figure 10 presents the same data points but now colored by language. For many of the semantic categories, the English words tend to group together near the centroids of their respective categories. On the one hand, this might be due to the English embeddings having a lower variance than the Mandarin embeddings.[4] On the other hand, closer inspection of individual clusters suggests that the procrustes analysis points to some non-trivial within-category differences between Mandarin and English.

In what follows, we zoom in on three categories in Figure 10: PERSON, FOOD, and BODY. Figure 11 shows a scatterplot of the nouns in the PERSON category. The color coding differentiates between person nouns denoting kinship (red), occupation (blue), and people (green). Here, we focus on the kinship terms. For Mandarin, these are found in the lower left (Group 1) and lower right (Group 2), whereas for English, these are found in a single cluster in the upper left (Group 3).

The kinship terms in Group 1 are mostly words denoting the closest family members (e.g. 爸爸 *bà-ba* "dad" and 妈妈 *mā-ma* "mom"). By contrast, words for more distant family members are predominant in Group 2. Some of these words are also used for polite referencing of non-relatives, similar to the use of brother and sister in English to refer to people from the same church. In Group 2, we also find words such as 阿姨 *ā-yí* "aunt, mother's sister", which is a polite form of address used for nannies and other caretakers in the home. The kinship terms of English form one cluster, in which the somewhat more formal terms (e.g. sibling, grandparent, grandchild) are found more to the right. We added two supplementary words，结婚 *jié-hūn* in Mandarin and *marry* in English to the plot, based on their t-SNE coordinates. These verbs are outliers with respect to the cluster of social verbs. Perhaps unsurprisingly, 结婚 occurs close to Group 1, and *marry* close to Group 3.

Figure 12 presents a scatterplot of the words in the FOOD category. The nouns referring to fruits and vegetable are labelled as "natural" and colored in red. Staples, shown in brown, are positioned close to fruits and vegetables. For English, one cluster is found at the (25,-24) with some words scattered among the words in the area labelled as Group 2, which contains mainly staple foods. For Mandarin, 燕麦 *yàn-mài* "oats", 大麦 *dà-mài* "barley", and 小麦 *xiǎo-mài* "wheat" are found in the upper central cluster. At (12.5, -21), words for various kinds of unprocessed rice (e.g. 大米 *dà-mǐ*, 小米 *xiǎo-mǐ*, and 糯米 *nuò-mǐ*) and flour （面粉 *miàn-fěn*)form a small separate cluster.

The words in Group 1 denote processed foods specific to Chinese cuisine, including various kinds of rice and noodles, as well as various kinds of organ meats. For example, 毛肚 *máo-dǔ*, the stomach of a cow, can be cooked in many different ways, especially as food in hotpot.

[4] We calculated for every individual embedding the variance of its values. For English, the mean of these variances was 0.0045, and for Mandarin, 0.0316 ($t(3042.8) = 113.01, p < 0.0001$).

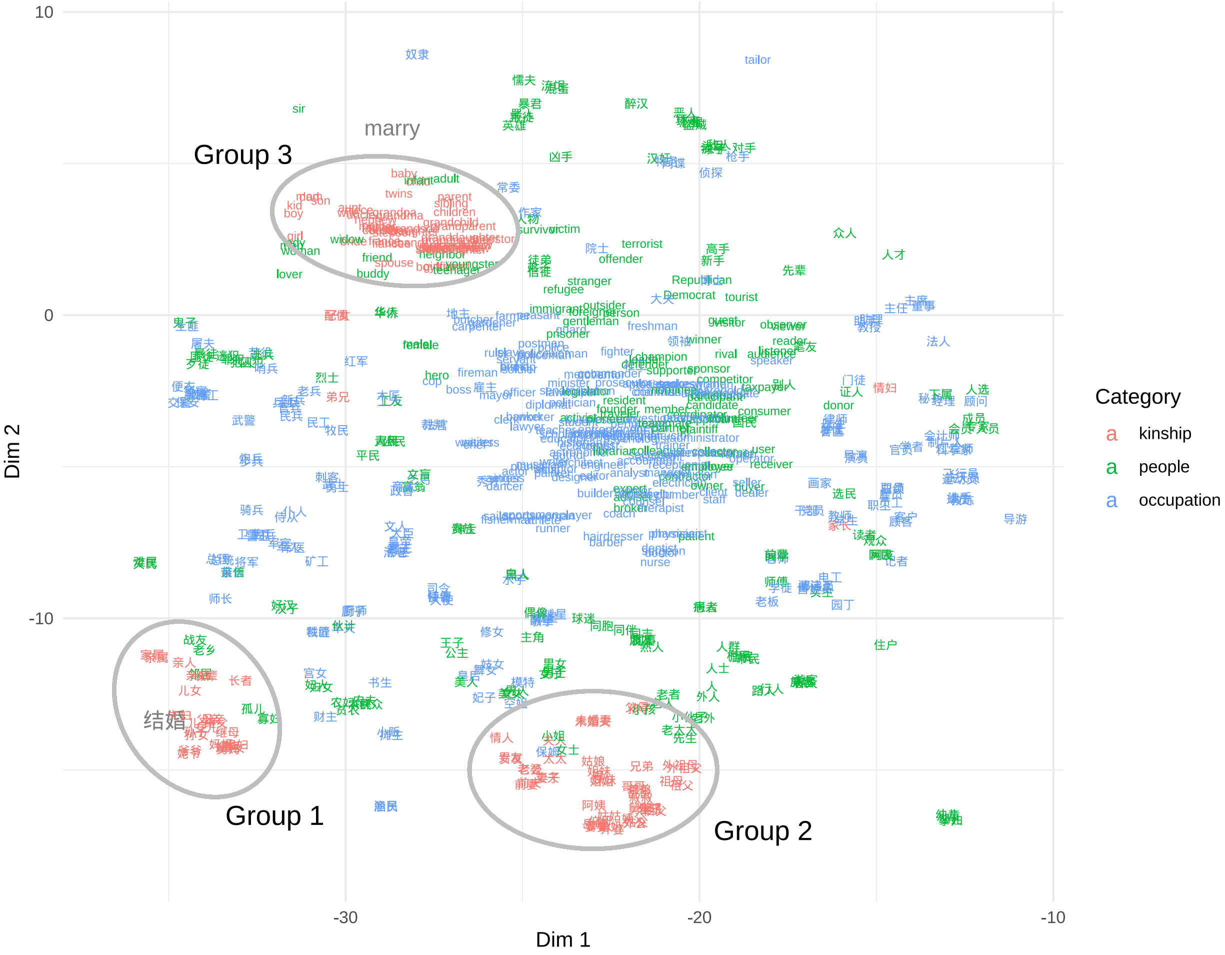

Figure 11: Zoomed-in t-SNE plot highlighting the distribution of person nouns.

The words in Group 2 are almost all from English, and denote various kinds of prepared foods. In the lower left of Group 2, we find meat products. Above these, words denoting composites of wheat and meat-based products (e.g. *burger* and *hotdog*). Further up, words denoting various kinds of wheat products without meat (e.g. *baguette*, *bread*, *brownies*), and at the top of Group 2, various diary products cluster together.

To the lower right of Group 2, a smaller cluster of Chinese processed foods is located (Group 3), with clearly meat-based products to the left(e.g. 荤菜 *hūn-cài*, “food with meat”), and primarily grain-based products to the right (e.g. 饺子 *jiǎo-zi* “dumpling” and 馅饼 *xiàn-bǐng* “pie”). The word for egg, 鸡蛋 *jī-dàn*, is in between the two groups, it is used in many of the dishes on its left and right in Group 3.

In Figure 12, the words for garlic are highlighted with a larger font size. In both languages, these words occur in the proximity of words for vegetables, unsurprisingly. In Chinese, most vegetable dishes are prepared with garlic. By contrast, in English, garlic is more like a spice that is added to some vegetable dishes, just as other spices such as chilli and rosemary. Thus, in English, garlic in the center of a group of herbs, positioned somewhat further away from the vegetables compared to Mandarin.

Finally, Figure 13 zooms in on the cluster of the BODY category. Words for parts of the torso and some general words such as *blood* and *muscle* are shown in blue. Words for parts of the head, and words for limbs and parts thereof, are depicted in red.

For English, words for part of the torso and terminology that is more specific to medical texts (e.g., *sinus, pelvis*), are found more to the right. Words for limbs and the head form two distinct clusters, which are located more to the left. Words for parts of head (*eye, chin, mouth*) are found in the upper left in a cluster labelled “Head”. Words for limbs and parts of limbs (*leg*, *arm*, and *elbow*) are found in the bottom center of the plot. Various words for larger parts of the torso (e.g., *neck, waist, chest*, and *breast*) form a bridge between the English Limb and Head clusters. For the Mandarin BODY terms, we have highlighted the clusters of words relating to the head and parts of the head and words relating to the limbs and parts thereof.

The words for limbs (e.g. 手心 *shǒu-xīn* “palm”, 手背 *shǒu-bèi* “back of hand”, 腿 *tuǐ* “leg”, and 胳膊 *gē-bo* “arm”) are positioned more to the left, whereas the words expressing facial features are found more to the right. The two clusters are closer together compared to the corresponding clusters in English. They are also surprisingly far removed from other related. For instance, whereas for English, *hair* and *tooth* are close to the Head cluster, the corresponding words in Mandarin, 头发 *tóu-fa* and 牙齿 *yá-chǐ* are located at a great distance in the upper right of the scatterplot.

In summary, we used a procrustes rotation to align the semantic spaces of Mandarin and English. This rotation is remarkable successful in aligning the semantic categories of the two languages. Fortunately, the procrustes alignment does not enforce complete alignment, and differences in semantic structure are also brought to the fore. At this stage of our research into the culture-specific aspects of semantic structure, all we can do is observe. The challenge for future research is to proceed observing and to also explain the observed differences.

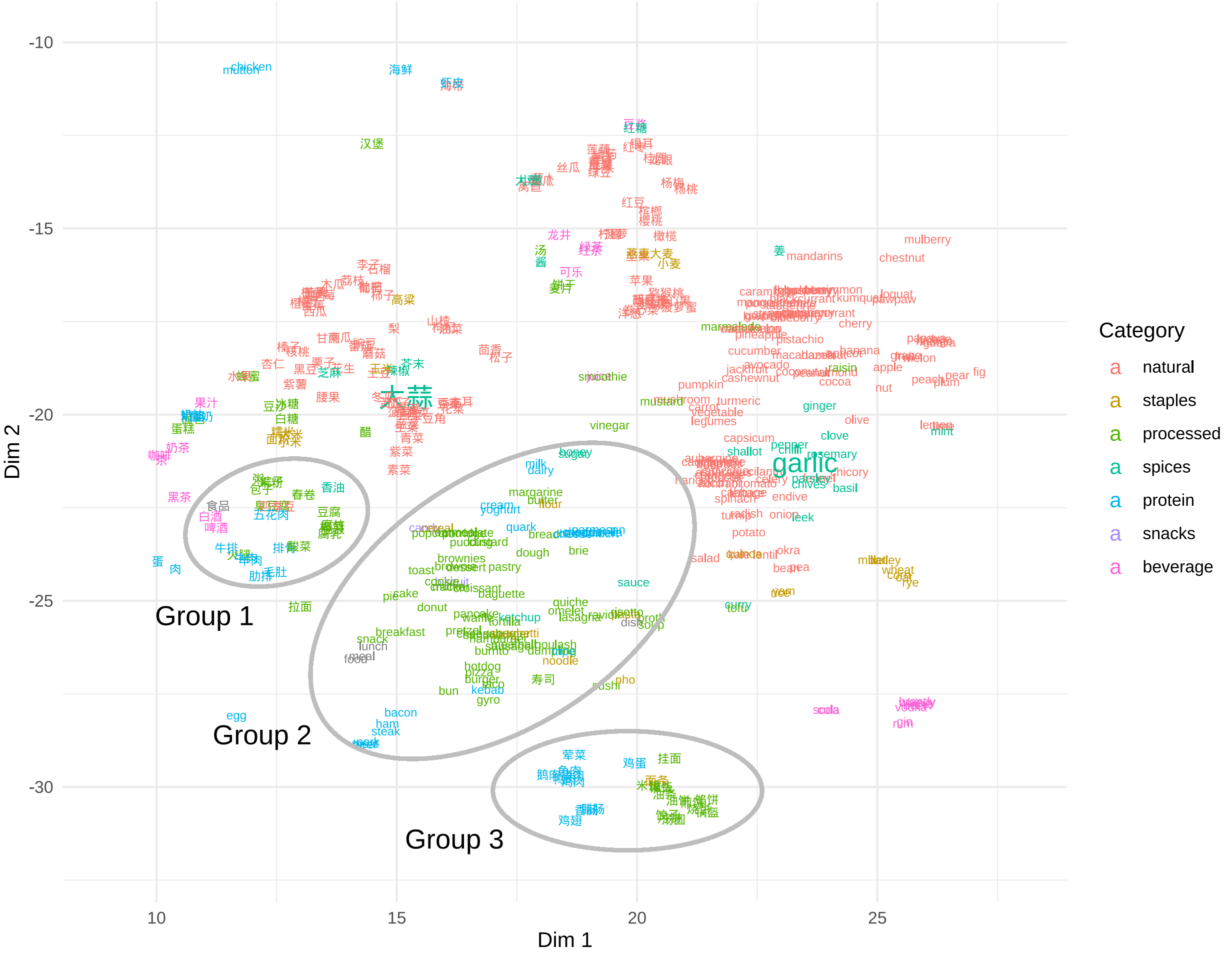

Figure 12: Zoomed-in t-SNE plot highlighting the distribution of FOOD nouns.

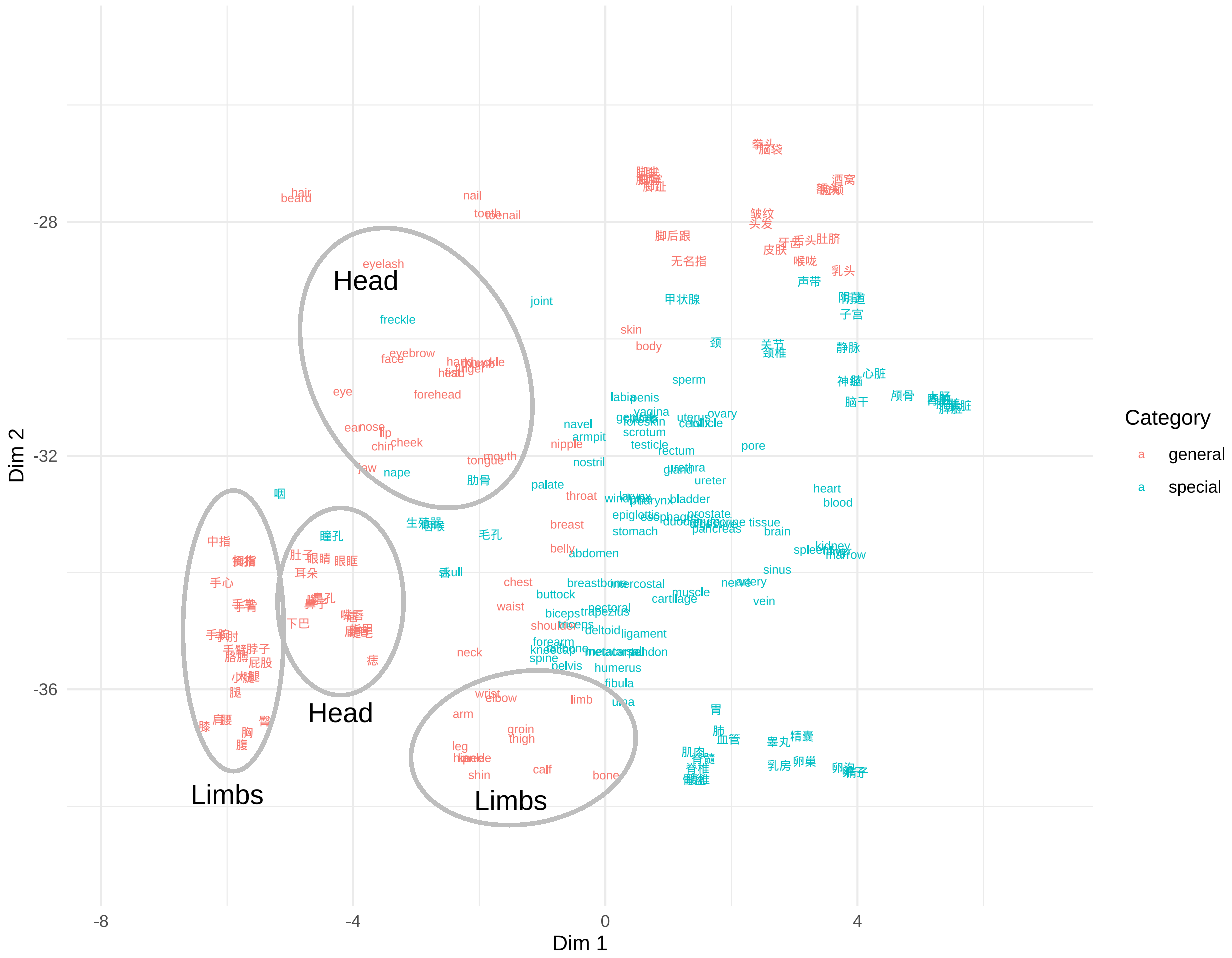

Figure 13: Zoomed-in t-SNE plot highlighting the distribution of BODY nouns.

6 General discussion

The present study investigated how words from different semantic categories cluster in the semantic spaces of Mandarin and English, using distributional semantics.

We first explored the two languages side by side using several techniques of dimensionality reduction. We also calculated by-category average vectors and explored the distances and cosine similarties between the semantic categories.

The semantic spaces of Mandarin and English have many features in common, as expected, given that despite geographical separation, language users, regardless of their native tongue, have many experiences in common. For instance, in both languages, nouns are segregated from other parts of speech such as adjectives and verbs. Evaluative adjectives and modals are closely associated with verbs. Nouns denoting natural entities form one subgroup, while those representing human-made entities constitute another. Time expressions form outlier clusters.

The semantic spaces of Mandarin and English also reveal some clear differences, which are highlighted most clearly by a graph-based analysis. In the Mandarin graph, the person category is associated with the one node that links a large cluster of nouns with a distinct second cluster of verbs and adjectives. In the English graph, by contrast, they occupy a relatively marginal position, with as closest neighbors the categories of supernatural beings and animals. This suggests that in Mandarin, the category of person is more integrated in the semantic system compared to English, perhaps reflecting a more collective understanding of persons as social agents in Mandarin as compared to English.

For English, the category of onomatopoeia, which comprises mostly verbs, is positioned close to the position and perception verb categories. For Mandarin, by contrast, the onomatopoeia behave more like adverbial expressions that describe the sound or force of actions. They are not closely linked to specific verbs, which may explain why in the graph of Mandarin, they are represented by an unconnected node. In other words, although Mandarin and English both have means of expressing sounds, the semantics of sound symbolism are remarkably different and also diverge considerably in how they are put to use in the syntax.

As a second step, we made use of procrustes analysis to compare two semantic systems in the same space. In the shared procrustean space, words cluster primarily cluster by semantic category, rather than by language, indicating that the procrustes rotation, which we defined on the basis of category centroids, is effective. Within the semantic clusters, subtle differences between the two languages emerge. For instance, a cluster of English words for processed foods is situated in between two clusters of Mandarin words, one of which brings together foods that are not part of English cuisine (e.g. 包子), and foods that are more similar to those that are part of English cuisine (e.g., 馅饼 *xiàn-bǐng* "pie"). The direct neighbors of "garlic"/ 大蒜 in Mandarin and English also diverge considerably.

One of the dimension reduction techniques that we used, t-SNE, has not been used extensively in previous corpus-linguistic studies (for examples, see Perek, 2018; Stupak and Baayen, 2022; Chuang et al., 2022). T-SNE outperformed PCA and MDS in finding clear clusterings. The much clearer clusters that emerge from the t-SNE are helpful for understanding the similarities between the different semantic categories. However, it is important to keep in mind that the relative distance between clusters in a t-SNE plot do not reflect distances in semantic space. PCA and MDS plots reveal similar category positions in a less distinct manner, highlighting an important property of semantic categories, namely, their fuzziness (Rosch and Mervis, 1975; Rosch, 1975).

A methodological innovation of the present study is the use of procrustes analysis to study Mandarin and English embeddings in the same high-dimensional space. As our English and Mandarin words do not form translation pairs, we calculated the procrustes rotation on the basis of the paired category centroids, and then applied this rotation to all words. The procrustes analysis turned out to be surprisingly effective in aligning two semantic spaces, making it possible to compare word embeddings from different languages (and different semantic spaces).

We deliberately avoided using multilingual transformers. Multilingual transformers such as developed by Xue et al. (2020) and Workshop et al. (2022) are trained on large numbers of languages, including programming languages, with most training materials coming from English and other western Indo-European languages. As shown by Wendler et al. (2024), such transformers have an English bias. Even in the unlikely case that multilingual transformers would be trained on the same amounts of data from languages balanced for language family, the result would be an "artificial universal speaker" that is a balanced blend of all languages sampled, yet unfaithful to any individual language when it comes to the details. Our interest, by contrast, is in the fine details in which the conceptual systems of languages differ.

The present study also has several limitations. First, the number of words that we included is small compared to the vastness of Mandarin and English vocabularies. Second, we assigned a word to one category only, simplifying the true complexities of categories and their overlap. Third, the selection of words included in the different categories has a subjective component. Our categories are therefore tentative, and likely to have language-specific cultural biases. Fourth, the FastText embeddings that we used are likely to represent blends of words' actual context-specific senses (Desagulier, 2019). We assigned words to categories based on their dominant sense, but even so, these dominant senses are not represented in semantic space with the precision that we would have liked to have. On the other hand, the fact that the embeddings for 树 *shù* and *tree* are dominated by the concept of natural trees, and are hardly influenced by the use of the word *tree* in linguistics to refer to particular kind of graphs, is perhaps a good thing. Fifth, the Fasttext embeddings for Mandarin were trained on corpora written in both simplified and traditional Chinese. We find a clear and strong distinction between words only used in simplified Chinese and words used in both writing systems. Interestingly, this distinction dominates in one dimension only and does not play a substantial role in other PCA or MDS dimensions. Furthermore, we have not been able to relate differences on the orthographic dimension to differences in meaning. This indicates that it is unlikely that the results of the present study are qualitatively affected by the way in which words are written in Mandarin.

In corpus linguistics, word embeddings have a long history of use. Baayen and Moscoso del Prado Martín (2005) used embeddings to study differences in the semantics of regular and irregular verbs in English, German and Dutch. Embeddings have been found useful for studying semantic change over time (Hilpert, 2014), for clarifying word senses (Hilpert and Flach, 2020), and for addressing conjectures about asymmetric priming (Hilpert and Saavedra, 2020). Embeddings have also been found to be informative for the study of semantic transparency in morphology (Marelli and Baroni, 2015; Shen and Baayen, 2022; Denistia et al., 2022) and the semantics of nominal pluralization (Shafaei-Bajestan et al., 2024). The main contributions of the present exploratory study to this growing body of literature is to show how embeddings can be used to trace the structure of the lexicon as a semantic system comprising many different semantic classes, and to pave the way for comparing the semantic systems of other different languages and cultures, such as French, Estonian, Persian, Hindi, Japanese, Arabic, and Korean. Word embeddings for these languages (and many others) are available on FastText. We hope that the present approach, and procrustes analyses using the group centroids of semantic categories, will be found useful for these languages as well.

References

Arora, S., May, A., Zhang, J., and Ré, C. (2020). Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*.

Asgari, E. and Schütze, H. (2017). Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Baayen, R. H. and Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, 81:666–698.

Baayen, R. H., Van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Black, P. (1973). Multidimensional scaling applied to linguistic relationships. *Cahiers de l'Institut de Linguistique de Louvain*, 3:n5–6.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017a). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017b). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Chen, A. C.-H. (2022). Words, constructions and corpora: Network representations of constructional semantics for mandarin space particles. *Corpus Linguistics and Linguistic Theory*, 18(2):209–235.

Chuang, Y.-Y., Brown, D., Baayen, R. H., and Evans, R. (2022). Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon*, 17(3).

Cox, M. A. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg.

Croft, W. and Poole, K. T. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics*, 34(1):1–37.

Davies, M. and Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. Routledge.

Denistia, K., Shafaei-Bajestan, E., and Baayen, R. H. (2022). Exploring semantic differences between the indonesian prefixes pe- and pen- using a vector space model. *Corpus Linguistics and Linguistic Theory*, 18(3):573–598.

der Klis, M. V. and Tellings, J. (2022). Generating semantic maps through multidimensional scaling: linguistic applications and theory. *Corpus Linguistics and Linguistic Theory*, 18(3):627–665.

Desagulier, G. (2019). Can word vectors help corpus linguists? *Studia Neophilologica*, 91(2):219–240.

Dictionary Office, Institute of Linguistics, C. (2016). *Modern Chinese Dictionary (7th Edition)*. The Commercial Press.

Divjak, D. and Gries, S. (2009). Behavioral profiles.: A corpus-based approach to cognitive semantic analysis. In *New directions in Cognitive Linguistics*, pages 57–75. John Benjamins Publishing Company.

Fox, R. A., Flege, J. E., and Munro, M. J. (1995). The perception of english and spanish vowels by native english and spanish listeners: A multidimensional scaling analysis. *The Journal of the Acoustical Society of America*, 97(4):2540–2551.

Gandour, J. T. and Harshman, R. A. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and speech*, 21(1):1–33.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh University Press.

Hilpert, M. and Flach, S. (2020). Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities*, 36(2):307–321.

Hilpert, M. and Saavedra, D. C. (2020). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*, 16(2):393–424.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

Laakso, A. and Smith, L. B. (2007). Pronouns and verbs in adult speech to children: A corpus analysis. *Journal of Child Language*, 34(4):725–763.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Levshina, N. (2015). *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. John Benjamins Publishing Company.

Levshina, N. (2016). Verbs of letting in germanic and romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast*, 16(1):84–117.

Marelli, M. and Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

Musil, T. (2019). Examining structure of word embeddings with pca. In *International Conference on Text, Speech, and Dialogue*, pages 211–223. Springer.

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H. B. A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M. O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C. J., and Weedon, J. (2022). *vegan: Community Ecology Package*. R package version 2.6-4.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559—572.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):65–97.

Peres-Neto, P. R. and Jackson, D. A. (2001). How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the mantel test. *Oecologia*, 129:169–178.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R. H. (2024). The pluralization palette: Unveiling semantic clusters in english nominal pluralization through distributional semantics. *Morphology*, page accepted for publication.

Shaoul, C. and Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavior Research Methods*, 42(2):393–413.

Shen, T. and Baayen, H. (2023). Productivity and semantic transparency: An exploration of word formation in mandarin chinese. *The Mental Lexicon*.

Shen, T. and Baayen, R. H. (2022). Adjective–noun compounds in mandarin: a study on productivity. *Corpus Linguistics and Linguistic Theory*, 18(3):543–572.

Stupak, I. and Baayen, R. H. (2022). An inquiry into the semantic transparency and productivity of german particle verbs and derivational affixation. *The Mental Lexicon*, 17(3).

Stupak, I. V. and Baayen, R. H. (2023). An inquiry into the semantic transparency and productivity of german particle verbs and derivational affixation. *The Mental Lexicon*.

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

Venables, W. N. and Ripley, B. D. (2002). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R Foundation for Statistical Computing, Vienna, Austria.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Wendler, C., Veselovsky, V., Monea, G., and West, R. (2024). Do llamas work in English? On the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

White, A. S., Hacquard, V., and Lidz, J. (2018). Semantic information and the syntax of propositional attitude verbs. *Cognitive Science*, 42(2):416–456.

Wilkes, A. (2008). *Chinese-English Visual Bilingual Dictionary*. Dorling Kindersley Ltd.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xiao, R., Rayson, P., and McEnery, T. (2015). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. Routledge.

Xu, X., Li, J., and Chen, H. (2022). Valence and arousal ratings for 11,310 simplified chinese words. *Behavior research methods*, 54(1):26–41.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Words for the category of BODY

• Mandarin words in the category of BODY

背	鼻孔	鼻子	脖子	唇	大肠	大腿	肚脐
肚子	额头	耳朵	肺	腹	肝脏	睾丸	胳膊
骨	骨盆	关节	喉咙	肌肉	脊髓	脊椎	甲状腺
肩	脚	脚后跟	脚尖	脚腕	脚掌	脚趾	睫毛
精囊	精子	颈	颈椎	静脉	酒窝	肋骨	脸
脸颊	颅骨	卵巢	卵泡	卵子	毛孔	眉毛	拇指
脑	脑袋	脑干	内脏	皮肤	脾脏	屁股	拳头
乳房	乳头	舌	舌头	神经	肾脏	生殖器	声带
食指	手	手背	手臂	手腕	手心	手掌	手肘
瞳孔	头	头发	腿	臀	胃	无名指	膝
下巴	小腿	心脏	胸	血管	牙齿	咽	咽喉
眼睛	眼眶	腰	腰椎	胰脏	阴道	阴茎	指甲
痣	中指	皱纹	子宫	嘴	嘴唇		

• English words in the category of BODY

abdomen	ankle	appendix	arm	armpit	artery	back	beard
belly	biceps	bladder	blood	body	bone	brain	breast
breastbone	buttock	calf	cartilage	cervix	cheek	chest	chin
clitoris	deltoid	digestive	duodenum	ear	elbow	endocrine	epiglottis
esophagus	eye	eyebrow	eyelash	face	fibula	finger	fist
follicle	foot	forearm	forehead	foreskin	freckle	genitals	gland
groin	hair	hand	head	heart	heel	hip	humerus
instep	intercostal	jaw	joint	kidney	knee	kneecap	knuckle
labia	larynx	leg	ligament	limb	lip	liver	lung
marrow	metacarpal	metatarsal	mouth	muscle	nail	nape	navel
neck	nerve	nipple	nose	nostril	ovary	palate	pancreas
pectoral	pelvis	penis	pharynx	pore	prostate	rectum	scrotum
shin	shoulder	sinus	skeleton	skin	skull	sperm	spine
spleen	stomach	tailbone	tendon	testicle	thigh	throat	thumb
tissue	toe	toenail	tongue	tooth	trapezius	triceps	ulna
ureter	urethra	uterus	vagina	vein	waist	windpipe	wrist

B T-SNE, UMAP, and PaCMAP

UMAP (McInnes et al., 2018) and PaCMAP (Wang et al., 2021) provide alternative clustering methods that have been argued to be superior to t-SNE. For instance, Wang et al. (2021) show that PaCMAP better preserves or reconstructs the topology of a three-dimensional dataset compared to both t-SNE and UMAP. Figure 14 presents scatterplots for the Mandarin data using t-SNE (top panel), UMAP (center panel), and PaCMAP (lower panel). For our data, t-SNE produces more distinct clusters than UMAP or PaCMAP. The relative positions of clusters are fairly similar for t-SNE and UMAP, and differ remarkably for PaCMAP. The PaCMAP is much more sensitive to whether words have distinct counterparts in traditional Chinese. This distinction is very strong on the first dimension of the PaCMAP and is also a strong separator on dimension 2. As it is unclear to us to what extent the similarity structure in the high-dimensional spaces that we are dealing with can be preserved with a model that has been evaluated on three-dimensional examples, we are unsure what to make of the output of PaCMAP, which we find less straightforwardly interpretable. In the light of these considerations, we conclude that t-SNE is an excellent choice for the purpose of our study.

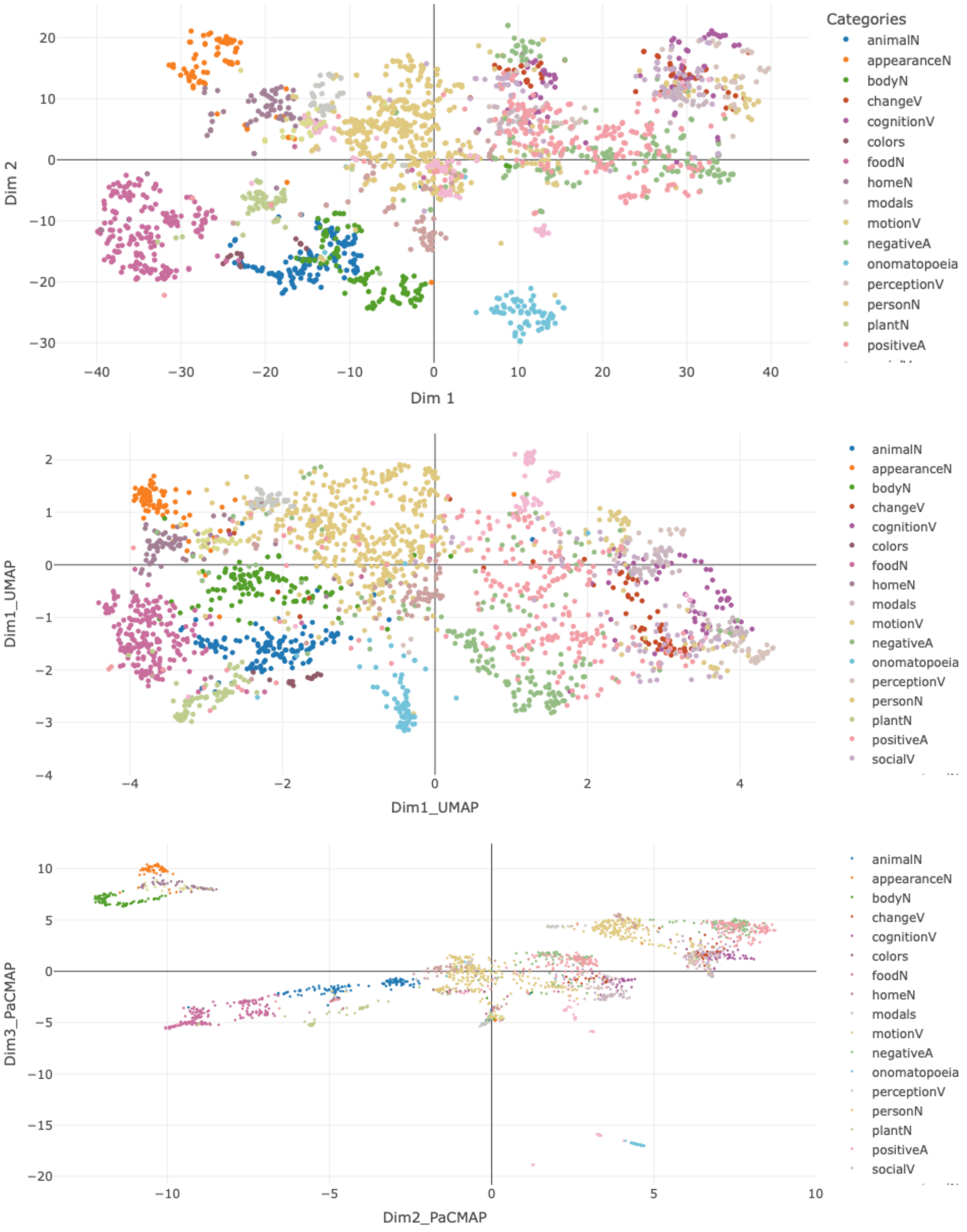

Figure 14: A comparison of t-SNE (top panel), UMAP(center panel), and PaCMAP (lower panel) clusterings of Mandarin words.