Treat the Word As a Whole or Look Inside? Subword Embeddings Model Language Change and Typology

Yang Xu
Department of Computer Science
San Diego State University
yxu4@sdsu.edu

Jiasheng Zhang
College of Info Sci and Tech
The Pennsylvania State University
jpz5181@ist.psu.edu

David Reitter
Google AI
reitter@google.com

Abstract

We use a variant of word embedding model that incorporates subword information to characterize the degree of compositionality in lexical semantics. Our models reveal some interesting yet contrastive patterns of long-term change in multiple languages: Indo-European languages put *more* weight on subword units in newer words, while conversely Chinese puts *less* weights on the subwords, but more weight on the word as a whole. Our method provides novel evidence and methodology that enriches existing theories in evolutionary linguistics. The resulting word vectors also has decent performance in NLP-related tasks.

1 Introduction

The roles that subword units play in determining word semantics differ across languages. In typical alphabetic languages, such as English, the smallest grammatical subword unit is *morpheme* (Katamba, 2015). A morpheme can be classified as either free or bound: the former stands by itself as a word (e.g., the *root* of English words), while the latter functions only as part of a word (e.g., *affixes* such as *-ness*, *un-*, etc.). In Eastern-Asian languages, however, the distinction between morphemes and words is not as clear. Particularly in Chinese, the basic subword unit that acts as a morpheme is character (字), but whether a single morpheme or the combination of morphemes constitute a word is open to debate (Hsieh, 2016).

Despite the fact that morphological regularities of words have been extensively applied to improve the dense vector representations of words learned from data, i.e., word embeddings (Chen et al., 2015; Bojanowski et al., 2017; Xu et al., 2018b), the research endeavors so far are less oriented towards linguistic theories about the semantic roles of subword units in word formation. In other words, NLP research has optimized towards processing languages such as English, but less so Chinese.

This study provides a first attempt (to the best of our knowledge) that uses word embedding models to explore the roles of subword units in the composition of word meanings. The source code is available at https://github.com/innerfirexy/lchange2019. We have shown that a variant based on the current subword-incorporated models can effectively quantify the semantic weights carried by subword units, with the cost of an moderate number of additional parameters, which have clear interpretations. Moreover, we have found that these semantic weights demonstrate temporal patterns that are different between Chinese and Indo-European languages, which implies a fundamental difference in the mechanisms of word formation. More theoretical motivations are discussed in the following section.

1.1 Theoretical Motivation

The direct motivation of this study is based on an empirical conclusion about the evolution of the Chinese language: the relative predominance of the *monosyllabic* words (i.e., single character as a word) in ancient Chinese has shifted to *bisyllabic* words in modern Chinese (Hsieh, 2016), and the long-existing yet unresolved disagreement regarding what a word is in Chinese among lay speakers and linguists (Sproat and Shih, 1996). In other languages, variationist inquiry has turned up regular patterns of shifting from a synthetic (single-word) to analytic (multi-word) constructions. Examples include *des Hauses* (the house's)→*von dem Haus* (of the house), *Edith chanta* (Edith sang)→*Edith a chanté* (Edith has sung) (Haspelmath and Michaelis, 2017). Though these observations are at the phrase level, it is reasonable to check if similar patterns can be found at the word level, because of the self-similarity property in natural language (Shanon, 1993).

The recently developed techniques of learning vector representations of words from data provide a new angle to revisit and contemplate the above theoretical confusions. For example, if we manage to quantify the semantic weight that a Chinese character carries in a word, then we can use it to verify the hypothesis that individual characters play less role in modern Chinese, which is more fine-grained evidence than mere frequency-based statistics.

In this study, we propose an approach that fits additional, theoretically informative parameters to configure a mixture of embeddings. With this, we characterize the relative contributions from words and subword units and capture them with an embedding model.

2 Related Work

2.1 Learning vector representations of words

Among the massive amount of work on learning dense word vectors, one of the most popular method is the word2vec model, which implements two efficient ways of learning word vectors, skipgram and CBOW (continuous bag of words) (Mikolov et al., 2013b,a). Both models learn word embeddings by training a network to predict words that co-occur within a window.

CBOW aims at predicting the target word given context words in a fixed window. For a training dataset of size T from a corpus of vocabulary size V, the learning objective of CBOW is to maximize the log probability: $L_{\text{CBOW}} = \sum_{i=1}^{T} \log p(w_i|C_i)$, where w_i is the target word, and C_i represents the surrounding context words The probability $p(w_i|C_i)$ is formulated by a softmax function:

$$p(w_i|C_i) = \frac{\exp(u_i^\intercal \cdot \boldsymbol{v}_c)}{\sum_{j \in V} \exp(u_j^\intercal \cdot \boldsymbol{v}_c)} \tag{1}$$

$$\text{in which } \boldsymbol{v}_c = \frac{1}{|C_i|} \sum_{w_k \in C_i} \boldsymbol{v}_k \tag{2}$$

where $\boldsymbol{v}_c$ is the average vector of all context words, and $\boldsymbol{v}_k$ is the vector of kth context word w_k, and u_i is the vector of the target word.

Skipgram predicts the context word given the target word at the center, by maximizing the log probability objective: $L_{\text{SG}} = \sum_{i=1}^{T} \sum_{w_k \in C_i} \log p(w_k|w_i)$, in which the probability $p(w_k|w_i)$ is also derived from a softmax function:

$$p(w_k|w_i) = \frac{\exp(\boldsymbol{v}_i^\intercal \cdot u_k)}{\sum_{j \in V} \exp(u_j^\intercal \cdot \boldsymbol{v}_i)} \tag{3}$$

where u_k is the context word and $\boldsymbol{v}_i$ is the target word. Because the softmax function is impractical to use due to its large amount of computation, hierarchical softmax or negative sampling are used when training the models (Mikolov et al., 2013b,a).

2.2 Word embeddings with subword information

For most languages in the world, the internal structure of words contain information about the semantics of the word. Incorporating parameters associated with those internal structures in the training process can improve word embeddings so that they are more expressive of the meanings of words. We deem that the improvements come from two sources: *semantic compositionality* and *reducing sparsity*.

Improvement from semantic compositionality
Some languages have strong compositionality at the word level. A good example is Chinese, of which a word is usually composed of several characters, and the meaning of the word can be inferred by assembling the meanings of all characters. For instance, the word “教育” (education), can be inferred from the meanings of its first character “教” (teach) and second character “育” (raise). Based on this thought, Chen et al. (2015) propose a character-enhanced word embedding model (CWE) that replaces the context word vector, $\boldsymbol{v}_k$ in eq. (1), with an average vector $\boldsymbol{x}_k$,

$$\boldsymbol{x}_k = \frac{1}{2}\boldsymbol{v}_k + \frac{1}{2}\Big(\frac{1}{N_k}\sum_{t=1}^{N_k} \boldsymbol{c}_t\Big) \tag{4}$$

where N_k is the number of characters in word w_k, and $\boldsymbol{c}_t$ is the vector of the tth character. Here the weights on the word and the characters within that word are equal (0.5), which is based on an empirical hypothesis that context words and characters are equally important to determine the semantics of target word. This is an over-simplicity that is reconsidered in our proposal.

Improvement from reducing sparsity
In some morphologically rich languages, one word can have multiple forms that occur rarely, making

it difficult to learn good representations for them. For example, Finnish has 15 cases for nouns[1], while French or Spanish have more than 40 different inflected forms for most verbs. A way to deal with this sparsity issue is to use subword information. Bojanowski et al. (2017) propose to learn representations for character n-grams and represent words as the sum of their n-gram vectors.[2] Their model, *fastText*, alters the training objective of skipgram by replacing the target word vector v_i with the sum of its n-gram vectors. Taking the word *love* for instance, it is represented by the following n-grams ($n = 3$): `<lo, lov, ove, ve>`, in which $<$ and $>$ are special symbols indicating the beginning and end of words. Each of these trigrams is associated with its own vector. Then the vector of *love*, $\vec{v}_{\texttt{love}}$, is computed as $\vec{v}_{\texttt{love}} + \vec{v}_{\texttt{<lo}} + \vec{v}_{\texttt{lov}} + \vec{v}_{\texttt{ove}} + \vec{v}_{\texttt{ve>}}$, i.e., the summation of all ngram vectors plus the word vector itself. More generally, fastText replaces the target word vector v_i in skipgram (eq. (3)) with an average vector x_i,

$$\boldsymbol{x}_i = \boldsymbol{v}_i + \sum_{t=1}^{N_i} \boldsymbol{c}_t \tag{5}$$

where N_i is the number of n-grams in word w_i, and $\boldsymbol{c}_t$ is the vector of the tth n-grams. The ideas of fastText and CWE are quite similar, only except that fastText is skipgram-based, while CWE is CBOW-based.

2.3 Word embeddings and language change

Word vectors have been used to study the long-term change of languages from multiple angles. The most straightforward method is to group text data into time bins and then train embeddings separately on these bins (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016). Conclusions about language change are reached by observing how the vectors of the same words change over time. The problem with this approach is that the learned word vectors are subject to random noise due to corpus size. Bamler and Mandt (2017) address this with a probabilistic variation of word2vec model, in which words are represented by latent trajectories in the vector space, and the semantic shift of words is described by a latent diffusion process through time.

Most of the existing approaches describe language change by the trajectories of some representations in a high dimensional space. Even though this provides rich information about every single point in the space (word, character etc.), it is difficult to interpret and summarize these models and discover the general patterns of language change.

3 Method

3.1 Dynamic subword-incorporated embedding model (DSE)

We propose the *Dynamic Subword-incorporated Embedding* (DSE) model, which captures the semantic weights carried by the subword units in words, on top of the architecture of CWE and fastText models. The "dynamic" part is reflected in a design considering that words rely on their internal structures to different degrees in composing a meaning: we associate each word in the vocabulary with a scalar parameter h^w, within the range $[0, 1]$, which is the weight of the word itself in predicting the co-occurred words within a context window. Correspondingly, $1 - h^w$ is the weight of its subword units. Here the subword units refer to characters in a Chinese word, and a subset of n-grams of a word for English and other four languages used in this study. We did not use word roots and affixes as the subword units as did by (Xu et al., 2018b), because of the lack of dictionary data in some languages, and the relative simplicity of n-gram-based models.

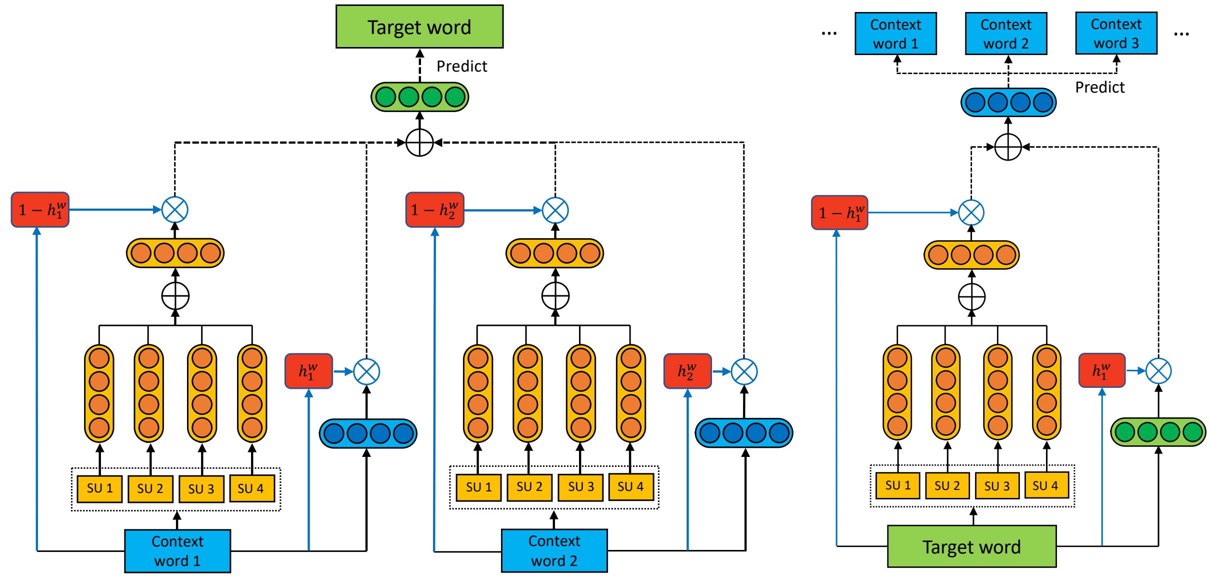
In DSE model, we use h^w to compute the weighted average vector for each word, and substitute it for the average context vector x_k in CWE model (eq. (4)), and for the average target vector x_i in fastText model (eq. (5)), as shown below:

$$\begin{cases} \boldsymbol{x}'_k = \boldsymbol{h}^{\boldsymbol{w}}_k v_k + (1 - \boldsymbol{h}^{\boldsymbol{w}}_k)\left(\frac{1}{N_k}\sum_{t=1}^{N_k} c_t\right), \\ \qquad \text{replacing the } x_k \text{ in eq. (4)} \\ \boldsymbol{x}'_i = \boldsymbol{h}^{\boldsymbol{w}}_i v_i + (1 - \boldsymbol{h}^{\boldsymbol{w}}_i)\sum_{t=1}^{N_i} c_t, \\ \qquad \text{replacing the } x_i \text{ in eq. (5)} \end{cases} \tag{6}$$

in which the subscripts k and i are the indices of words in the vocabulary. We have two versions of model architectures: one is based on CWE (CBOW-like), and the other is based on fastText (skipgram-like). They are referred to as *DSE-*

[1] See `http://jkorpela.fi/finnish-cases.html`

[2] Another approach is to tokenize words into subwords while optimizing a language model acquired over these *word pieces* (Schuster and Nakajima, 2012; Sennrich et al., 2015).



(a) DSE-CBOW

(b) DSE-SG

Figure 1: The architecture of the two version of DSE model. DSE-CBOW associates a semantic weight parameters h^w to each context word, and DSE-SG does this to each target word. The "SU"s in the yellow box stand for "subword units".

CBOW and *DSE-SG* respectively. The architectures of these models are shown in Figure 1.

We call h^w the semantic weight parameter. It describes the proportion of contribution from each word as a solitary semantic unit, while $1 - h^w$ is the total contribution from all the subword units. h^w is a learnable parameter in the model.

3.2 Corpus data and training setup

We use the Wikimedia database dumps[3] (up until July 2017) as our training data. Data in *six* languages are used: Chinese (ZH), English (EN), French (FR), German (DE), Italian (IT) and French (FR). Raw text data are extracted from the dump files using `WikiExtractor`[4]. Further text cleaning are conducted by separating sentences into per line, and converting non-proper-nouns (proper-nouns are identified using a pre-trained NER model provided in the Python package `spacy`[5]) to lower case. For Chinese data particularly, word segmentation is carried out using the `Jieba` segmenter[6]. All traditional Chinese characters are converted to simplified Chinese using OpenCC[7]. All non-Chinese characters are removed, keeping only those within the Unicode range U+4E00-9FFF. The training data of all six languages are of similar volumes: 33 to 40 million tokens each after preprocessing.

To accelerate training, we limit the number of *effective* semantic units in each word. For Chinese data, words containing more than 7 characters are ignored. For other languages, if a word contains more than 7 n-grams, we randomly select 7 out of them, and ignore the rest. Here the number 7 is chosen based on the following empirical observation: in a pilot study, we found that numbers larger than 7 will not improve the resulting embeddings, but significantly slow down the training. Other hyper-parameters are kept as close to the previous studies as possible. The detailed hyper-parameter settings and training procedures are described in Appendix A.

As for the size of n-grams, we use a fixed size $n = 4$, i.e., no bigrams or trigrams are considered. This choice is partially based on Bojanowski et al.'s (2017) work showing that $n = 4$ already achieves a satisfactory embedings, and partially due to speed consideration. For words that consist of than 4 letters, we only consider two sources for the mixture embeddings: the word itself and the n-gram ($n < 4$).

The semantic weight parameters h^w are implemented as a $V_w \times 1$ lookup table. Thus, in each training step, the learning algorithm updates three

[3] https://dumps.wikimedia.org/

[4] https://github.com/attardi/wikiextractor

[5] https://spacy.io/

[6] https://github.com/fxsjy/jieba

[7] https://github.com/BYVoid/OpenCC

embedding tables: word embeddings E_w, character embeddings E_c, and the semantic weights. Specifically, for DSE-SG model, the average embeddings are first computed from E_w, E_c, h^w, and h^c using eq. (6) and then outputted as the final word vectors. For DSE-CBOW model, just the E_w table is outputted as the learned word vectors[8].

4 Results and Discussion

4.1 Correlation between semantic weights and word ages

We are interested in examining the relationship between the semantic weight h^w of a word and its relative "age". According to the observation that Chinese is shifting from monosyllabic words to bisyllabic words, it is reasonable to expect that newer Chinese words should have larger h^w than those older words, because a higher h^w indicates that the word as a whole rather than the individual subword units is more important in determining its meaning. For other languages, we do not have a clear clue on what the relationship could be, but they should provide an interesting comparison.

First, we need to have a reliable way to measure the "age" of a word. We use the Google Books Ngram (GBN)[9] corpus, which contains word frequency information from about 10 million books published over a period of five centuries (Lin et al., 2012). It is the best resource we can find that provides estimated temporal distributions of words in multiple languages. For each word in GBN we extract the first *year* that it appears in the dataset, and use this first-appearance-year as an approximation of the word's age. Then we check if the word's age is correlated with its h^w from training the DSE model. For example, the word "爱人" (lover) first appears in the year of 1804 (AD) (at least according to the GBN collection). Thus, our examination is focused on the intersection of vocabularies between GBN and the training data. For DE, EN, ES, FR and IT, the intersection covers above 95% of the most common words in the training, and the proportion for ZH is 84%.

In a short summary of the results, we find *opposite* $h^2 \sim$ year relationships in Chinese and the other five languages. h^w *decreases* with the first-appearance-year in the five Indo-European languages, as shown in Figure 2. Words with subword units count ranging from 2 to 7 are included. Short words that have only 1 n-gram are excluded because the n-grams have the same form as the words. There are some fluctuations but the overall decreasing trends of h^w are salient. As the decrease of h^w is equivalent to the increase of $1-h^w$, it indicates that in these five languages, subword units carry more semantic weights in newer words than older ones. The h^w scores reported in Figure 2 are from DSE-SG, because those from DSE-CBOW are either 0s or 1s, due to the quick saturation of the softmax function, which however does not happen to Chinese data.

As for Chinese, however, h^w *increases* with the first-appearance-year as shown in Figure 3. We choose the subword units (characters) count = $\{2, 3, 4\}$ because they are the majority in the training data, with proportions 57.5%, 31.0%, and 8.6%. Frequency wise, their proportions are more dominant: 82.9%, 11.8%, and 4.6% respectively. Words composed of more than 4 characters are very uncommon in Chinese. From the plot, the increasing trends of the 2-character words are observable, but less so for the 3- and 4-character words. It indicates that our hypothesis in Section 1.1 is supported: characters carry more semantic weight in older Chinese words than in newer Chinese words.

Besides, an interesting finding is that the h^ws from DSE-SG are larger than those from DSE-CBOW in Chinese. It makes sense intuitively: a CBOW-like model is using multiple context words to predict one word, and thus the semantic weight from each individual word is diluted.

4.2 Statistical analysis to verify the results

Considering the fact that word frequency plays a critical role in a broad range of phenomena in language production and comprehension, including word length (Zipf, 1949), syntactic choice (Jaeger, 2010), and alignment (Xu et al., 2018a; Xu and Reitter, 2018), and the fact that during the training of a word embedding model, the more frequent words naturally get more updates on their embedding parameters, it is therefore necessary to rule out the possible confounding effect from word frequencies in the "$h^w \sim$ year" correlation found in previous section.

First, we fit a linear model with h^w as the response variable, and two predictors, the first ap-

[8] The discrepancy exists in the original implementations of CWE and fastText, and the reason behind is out of the scope of this study.

[9] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

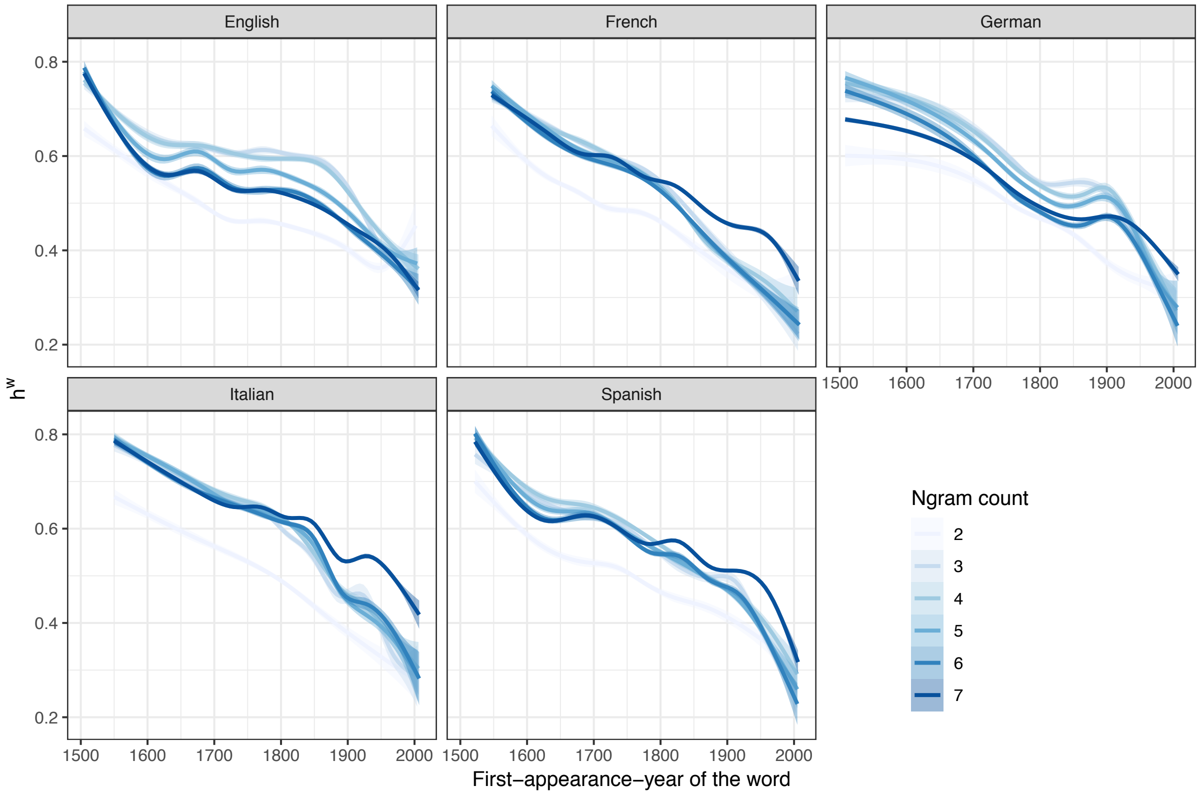


Figure 2: Semantic weight h^w against the first-appearance-year of words in DE, EN, ES, FR, and IT. Words with subword units (n-grams) number ranging from 2 to 7 are plotted separately. Shaded area indicates 95% point-wise confidence intervals of the fitted regression lines. h^w scores are from the DSE-SG model.

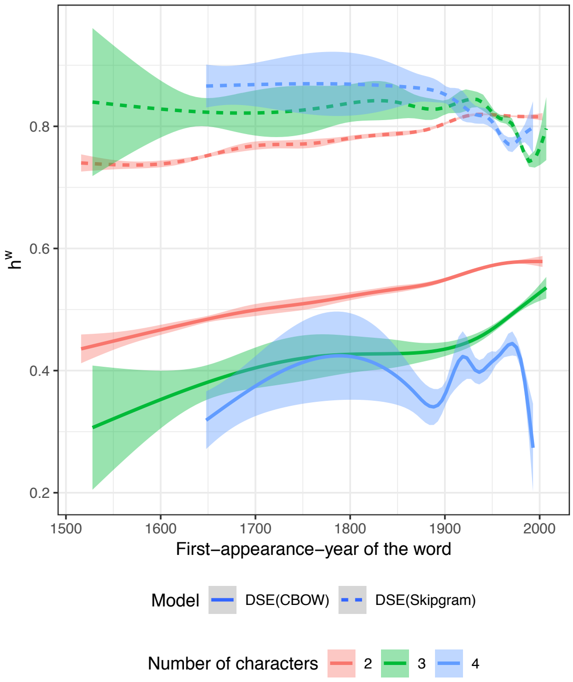


Figure 3: Semantic weight h^w against first-appearance-year for Chinese words with character number = 2, 3, and 4. Shaded area indicates 95% point-wise confidence intervals of the fitted regression lines.

pearance year and the frequency of words in training data, as expressed by the formula: $h^w \sim \text{year} + \text{frequency}$. We find that both covariates are significant predictors, as shown in the "Direct model" column of Table 1. The positive and statistically significant ($p < .001$) β_{year} coefficients indicate that the observed decreasing trend of h^w in five Indo-European languages and the increasing trend in Chinese are reliable, after the effect of word frequency is taken into account.

A more conservative method is to fit an auxiliary model m' first, with h^w as the response and word frequency as the sole predictor (regressing it out), and then fit a second model m, using the *residuals* of m' as the new response variable, and the first appearance year as its predictor. If the parameter estimate in m still indicates a significant effect, then that means the second predictor (year) indeed affects the response (h^w) in a way that is independent on the first predictor (frequency). With this step, we confirm the effect of year on h^w in all languages (see the "Auxiliary model" column in Table 1).

Language	β coefficient of year	
	Direct model	Auxiliary model
DE	$-7.8 \times 10^{-4***}$	$-7.8 \times 10^{-4***}$
EN	$-8.5 \times 10^{-4***}$	$-8.5 \times 10^{-4***}$
ES	$-8.5 \times 10^{-4***}$	$-8.4 \times 10^{-4***}$
FR	$-9.4 \times 10^{-4***}$	$-9.4 \times 10^{-4***}$
IT	$-9.2 \times 10^{-4***}$	$-9.1 \times 10^{-4***}$
ZH (DSE-SG)	$1.2 \times 10^{-4***}$	$1.0 \times 10^{-4***}$
ZH (DSE-CBOW)	$1.5 \times 10^{-4***}$	$1.3 \times 10^{-4***}$

Table 1: Statistical models to verify the decreasing trend of h^w with first-appearance-year in five Indo-European languages and its increasing trend in Chinese. *** indicates a significance level of $p < .001$.

4.3 Evaluation on lexical semantic tasks

Training the DSE model not just results in a lookup table of h^w, but also outputs word embeddings. In theory, these embeddings should be better representations of the semantic space than the CWE and fastText models, because DSE uses more parameters (h^w). Here, we compare the quality of word embeddings resulting from DSE models with those from previous models. Any superiority that DSE could show will indicate that dynamically considering the semantic weight of subword units can be potentially useful in other NLP tasks.

There are several standard lexical semantic tasks commonly used to evaluate the quality of embeddings. We use two of them, word similarity/relatedness and word analogy, and evaluate the embeddings from Chinese and English. For word similarity task, Wordsim-296 Chen et al. (2015) in Chinese and Wordsim-353 (Finkelstein et al., 2002) in English are used. Higher Spearman's correlation score indicates better performance. For word analogy task, the semantic part of the original dataset[10] developed by Mikolov et al. (2013a) and its Chinese translated version are used. The total percentage of correctly answered questions is used to measure the performance on this task.

The performance of DSE are shown in Table 2 compared with CWE and fastText. Here we do not use the original implementations of CWE and fastText (in C and C++), but use our own implementations with the same programming framework as DSE (By disabling the h^w parameters). This is for the consideration of fair comparisons. The models are compared within two groups according to the architecture: DSE-CBOW and CWE are in CBOW group; DSE-SG and fastText are in skip-gram group. We find that DSE-SG achieves *higher* score in word analogy task, and overall speaking, DSE models have comparable or slightly lower performance in word similarity task.

Language	Model	Similarity	Analogy
Chinese	DSE-CBOW	0.597	0.666
	CWE	0.605	0.668
	DSE-SG	0.583	**0.651**
	fastText	0.591	0.588
English	DSE-CBOW	0.659	0.302
	CWE	0.669	0.324
	DSE-SG	0.705	**0.356**
	fastText	0.702	0.338

Table 2: Performance in lexical semantic tasks.

It is surprising that DSE does not show a significant improvement, which could be due to the redundancy in model parameters or the size of training data. That said, what we show here is that the new model performs well enough to be plausible. We do not attempt to improve upon the state-of-the-art in these downstream tasks; rather, the main purpose of the model is for linguistic inquiries.

4.4 Case study

We use several cases of words to intuitively understand what specific aspects in language use causes the finding of this study. First, we find the magnitude of semantic weight h^w is related to the part-of-speech tag of words. For example, some earlier words that contain the character “安” (safe) are mostly used as adjectives, e.g., “安全” (secure, 1581), “安定” (settled, 1632) etc, while some newly appeared words are often nouns of terminology in certain fields, e.g., “安打” (*base hit*, a baseball term) first appears in 1959, “安检” (*security check*, an airport term) first appears in 1987. We find that the h^ws of the domain-specific nouns are higher than the generic adjectives (see Table 3), which indicates that the character “安” plays a lighter semantic role in these nouns. It also indicates that Chinese language users tend to consider the chunk of characters together as standalone semantic unit, and refer less to the original meanings of individual characters within.

We find similar cases in English. For example, the word *acid* first appears in 1517, and its h^w is larger than those words that contain the n-gram “acid” such as *acidosis* and *oxoacid*, both of

[10] http://download.tensorflow.org/data/questions-words.txt

Language	Older words	h^w	Newer words	h^w
Chinese	**安**全(secure), 1581 **安**定(settled), 1632	0.75 0.72	**安**打(base hit), 1959 **安**检(security check), 1987	0.85 0.87
	组成(consist of), 1568	0.67	课题**组** (research group), 1988	0.86
	覆**盖** (cover), 1747	0.69	**盖**帽(block)†, 1972	0.91
	把握(hold), 1591	0.69	拖**把** (mop), 1985	0.86
English	**acid**, 1517	0.73	**acid**osis, 1907 oxo**acid**s, 1953	0.07 0.07
	compare, 1524	0.86	**compar**ison, 1659 **compar**atives, 1810	0.61 0.14
	human, 1504	0.87	trans**human**ism, 1955	0.50
	locking, 1600	0.77	un**lock**able, 1854	0.11

†: A basketball term.

Table 3: Case study examples. Earlier words on the left are adjectives and verbs, and have smaller h^w. Later words on the right are nouns, and have larger h^w. Subword units shared across words are highlighted.

which first appear in 1900s. More examples are shown in Table 3. Similarly, some of the newer words (with higher h^w) are domain-specific compounds that consist of several seemingly unrelated n-grams, while some others are inflections of the original verb or adjectives.

We conjecture that a common cause for the changes of h^w in both languages can be the advancement of science and technology, and the need for new vocabulary that comes with it. However, the contrast between the two languages are intriguing: the relatively large h^w of new terms in Chinese seems to indicate that new meanings are assigned to these words without too much inference into the original meanings of the subword units; while the smaller h^w in English indicates the opposite, i.e., the original meanings of the containing subword units are emphasized more.

Through the examples, it indicates that the increasing trend of h^w may reflect the modernization of Chinese as the concepts and terminology in science and technology (and western culture as well) had been introduced since the 19th century, and more so ever after 1900s. Of course, the examples here do not exhaustively cover all possible causes for the change of h^w. We believe that an aggregative analysis over the effect of word types on h^w is necessary in order to get more comprehensive explanations.

5 Conclusions

In this study, we use a subword-incorporated model to characterize the semantic weights of subword units in the composition of word meanings. We find a major difference in the long-term temporal patterns of semantic weights between Chinese and five Indo-European languages: In Chinese, the weights on subword units (characters) shows a decreasing trend, i.e., individual characters play less semantic roles in newer words than older ones; In Indo-European languages, however, this trend is opposite, i.e., newer words place more weights on the subword units. In a more informal way: Chinese words are treated more as a whole semantic unit "synthetically", while words in Indo-European languages require more attention into the subword units "analytically".

Our findings provide new evidence to linguistic theories about word formation. First, the notion of "word" in Chinese is always changing: compared to its earlier age, modern Chinese tend to have multiple characters as a whole semantic unit. The semantic weight carried by a single character is decreasing. This is strong evidence in favor of the claim in qualitative studies that Chinese has been evolving towards multisyllabic from monosyllabic. Second, the increasing trend of semantic weights on subword units in Indo-European languages is consistent with the "synthetic → analytic" pattern shift at the phrase level composition (Hamilton et al., 2016). Moreover, the relative "synthetic" way of composing Chinese word found in this study seems consistent with the holistic encoding hypothesis in the perceptual theories about the Chinese writing system (Dehaene et al., 2005; Mo et al., 2015).

Going forward, we would like to apply the DSE model to other Eastern-Asian languages, such as

Korean and Japanese, which are not included in this study due to the lack of GBN data for them. If we find similar weighting patterns on subword units in these languages as in Chinese (which we anticipate to see), then we can have a bigger picture of the word formation mechanisms of different language families. Second, we would like to use roots and affixes instead of n-grams for Indo-European languages because their semantic meanings are more clearly defined, and thus knowing about their weights change with time can tell a better story of language evolution.

How the semantic meanings are conveyed in subword units is quite different in an Indo-European language such as English (where *word root*, *prefix*, and *postfix* play more important roles than characters) from the case of Chinese. However, it produces meaningful results to quantify and compare the semantic weights of subword units among different languages. At the core of this, there are questions of the universality of semantic subspaces across languages.

Acknowledgments

The authors thank Matthew A. Kelly, Frank E. Ritter, and Saranya Venkatraman for their comments. The authors also thank all the reviewers for their constructive feedbacks. The authors acknowledge support from the National Science Foundation, grant BCS-1734304 to D. Reitter.

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proc. 34th International Conference on Machine Learning*, pages 380–389.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *Proc. 24th International Joint Conference on Artificial Intelligence*, pages 1236–1242.

Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. 2005. The neural code for written words: a proposal. *Trends in Cognitive Sciences*, 9(7):335–341.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

M Haspelmath and SM Michaelis. 2017. Analytic and synthetic: Typological change in varieties of european languages. In *Language Vriation – European Perspectives VI. Selected Papers from the 8th International Conference on Language Variation in Europe*, pages 1–17.

Shu-Kai Hsieh. 2016. Chinese linguistics: Semantics. In Sin-Wai Chan, James W Minett, and Florence Li Wing Yee, editors, *The Routledge Encyclopedia of the Chinese Language*, pages 203–214. Routledge, Abingdon, Oxon.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Francis Katamba. 2015. *English Words: Structure, History, Usage*. Routledge.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Ce Mo, Mengxia Yu, Carol Seger, and Lei Mo. 2015. Holistic neural coding of chinese character forms in bilateral ventral visual system. *Brain and Language*, 141:28–34.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Benny Shanon. 1993. Fractal patterns in language. *New Ideas in Psychology*, 11(1):105–109.

Richard Sproat and Chilin Shih. 1996. A corpus-based analysis of mandarin nominal root compound. *Journal of East Asian Linguistics*, 5(1):49–71.

Yang Xu, Jeremy Cole, and David Reitter. 2018a. Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 601–610.

Yang Xu, Jiawei Liu, Wei Yang, and Liusheng Huang. 2018b. Incorporating latent meanings of morphological compositions to enhance word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1232–1242.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

A Appendices

The values of the hyper-parameters for training the DSE models are shown in Table 4.

Hyperparameter	Value
Embedding size	300 (Word) 300 (Subword)
Window size	5
Number of negative samples	10
Batch size	128
Minimal word frequency	5
Initial learning rate	0.05 (DSE-CBOW) 0.025 (DSE-SG)

Table 4: Hyperparameter settings.

The training stage consists of three steps:

- Pre-train the word embeddings: set the parameters for word embeddings, i.e., the v_k and v_i in Equation (6) trainable; set all the other parameters not trainable; train the model for 5 epochs.

- Pre-train the subword embeddings: set the parameters for subword units, i.e., c_t in Equation (6) trainable; set all the other parameters not trainable; train the model for 5 epochs.

- Set all the parameters trainable (including embeddings and h^ws); train the model for 5 epochs.