Quantifying semantic salience to investigate contact-induced language change

Online Proceedings of UK-CLA Meetings
http://uk-cla.org.uk/proceedings
Vol 4: 131 – 153

EILEEN WAEGEMAEKERS[1]

University of Hong Kong

eileen.waegemaekers@hku.hk

Abstract

Contact varieties that emerge in societies where multilingualism is the norm provide an ideal testing ground to study cognitive constraints on language change. One of these constraints is salience, which has been argued to be an important factor in determining whether a feature gets adopted or discarded in a contact variety (Aboh and Ansaldo 2006; Siemund and Kintana 2008). However, the notion of salience is problematic as it is an extremely broad concept causing scholars to each adopt slightly different definitions which are not always applicable cross-linguistically. In this paper, a new approach is put forward that quantifies semantic salience as the relative contribution of an individual word to the overall meaning of the sentence, coined as semantic load.

Models that apply principles of distributional and compositional semantics have shown success in capturing meanings of words (Erk 2012). Hence, a recursive neural network (RNN) is employed that incorporates these principles in its architecture, which is then trained on the task of missing word prediction (Le and Zuidema 2014a). Vector representations for words are based on their distributional properties in large text corpora and the sentence

[1] The author wishes to thank Phong Le, Sara Veldhoen and Jelle Zuidema (ILLC, University of Amsterdam) for assistance with the development of the computational model.

representations are formed during the task of word prediction. With these two types of vector representations it is possible to measure the semantic load of individual words.

This paper reports on semantic load as a cognitive constraint on language change by looking at the competition and selection of features with varying degrees of semantic load in data of contact languages. In doing so, it shows that computational models are not only useful in their domain of human-computer interaction but can also be employed as a tool in linguistic research. Moreover, as the neural network uses distributional and compositional representations of meaning, the results touch upon the cognitive plausibility of representing meaning in this manner.

Key words: semantic load, contact-induced change, neural network model, salience

1. Introduction

Language can be seen as a complex adaptive system (Kirby 1998; Steels 2000; Ellis & Larsen-Freeman 2009) that changes (i.e., "adapts") through usage. A question that many linguists have asked is whether it is possible to forecast in which direction a language will change and whether there are cognitive constraints that explain why languages change in a certain direction and not the other (Croft 2000; Matras 2009). Building on work by Croft (2000), Mufwene (2001), Ansaldo (2009) and Aboh (2015), it is assumed that language change at the individual level results from contact with different idiolects and that elements of those idiolects that the speaker has been in contact with, are in competition. That is, by interacting with different speakers with varying idiolects, either from the same language community or from a different language, the individual encounters variants for expressing the same thing. These variants will compete for selection. In a community, individual idiolects represent variations which are imperfectly replicated by language learners and users when selecting certain variants over others. A question that I wish to answer by the proposal put forward in this paper is whether inherent properties of the variants make one variant more likely to get selected over the other.

Mufwene (2001) coined the notion of feature pool from the analogy of 'gene pool' in biology. Features, in this approach, are the units and principles of a linguistic system (e.g., phonological, morphological and syntactic units or combinatorial and distributive principles) that are transmitted and selected in the feature pool. An individual speaker has a collection of features in her pool that are extracted from the input. With every interaction, features are introduced into the feature pool. When producing an utterance, the output is a

specific selection or recombination of features. Features in the feature pool can share some of their properties (e.g., they have a similar syntactic function, represent two variants with similar semantics or two different allophones). In the recombination process, competition arises when in the mind of an individual speaker there is an "unequal weighting of variants" (Mufwene 2008, pp. 116). The variant with the highest weight then gets selected. It is the unconscious selection of one feature over its variants that causes differential replication and can account for change in a language.

An important prerequisite for differential replication is competition between variants. The fact that speakers are able to describe the same idea in different ways is essentially the same as having two or more languages at your disposition to express that concept (see example (1), (2) and (3)) and results in competition.

1) I'm cycling

2) I'm riding my bike

3) Ik fiets ('I'm cycling' in Dutch)

A speaker of English can select either option (1) or (2) whereas a multilingual speaker of Dutch and English can select either (1), (2) or (3). Along these lines, the 'multilingual mind' can refer to any speaker. Even a strict monolingual speaker has variants in their language system. When a speaker receives input from different languages, the feature pool naturally contains features from different languages. Accordingly, in the process of recombination, the output may be a recombination of features from different languages. This recombinatory process of features in the mind of an individual speaker, both when it concerns idiolectal contact in a monolingual speaker and language contact in a multilingual speaker, is argued to be the primary cause of contact-induced language change.

At the communal level, it is the cumulative effect of individuals selecting one feature over the other or the hybrid blending of features that explains change over the course of time (Mufwene 2008). Features disappear from the feature pool when they are no longer selected by speakers in the speech community. At the individual level, features may disappear when they have not been selected for a long time (e.g., in the case of language attrition). The selections that are made by an individual speaker are unconscious and affected by both cognitive and social constraints. A question that is addressed in this paper is "what [cognitive] principles or constraints bear on such 'unconscious' selections into speakers' idiolects" (Mufwene 2008, pp. 120).

Studies in the field of language contact and the field of bilingualism and second language acquisition can shed light on what these constraints might be. By investigating what features are particularly vulnerable for change in

both multilingual speakers and situations of language contact, it is possible to reveal the constraints on the selection process. For example, Matras (2007) introduces the notion of borrowability, defined as "the likelihood that speakers will give up the separation between their 'languages' – the mental demarcation line that divides their overall repertoire of linguistic structures – in respect of a particular function-bearing structure ('category')" (p. 32). The borrowing hierarchies presented in Matras (2009) reveal "that the motivation to borrow is anchored in the intrinsic semantic-pragmatic function of the affected categories and in the contribution they make toward the mental processing of utterances in discourse" (pp. 163). For example, cross-linguistically 'but' is more often borrowed than 'or' and 'or' is more often borrowed than 'and'. He argues that this is the case because 'but' relates more to the presupposition domain than the other clause connectors as 'but' contrasts the interlocuters' shared presupposition. Moreover, it has been argued that features that have salient semantic or functional content are more likely to get selected in contact situations (Aboh and Ansaldo 2006; Matras 2011). An example from situations of intensive language contact, namely in areas where creoles developed, is the loss of verb inflections. In fact, it has been argued by McWhorter (2001) that lack of inflectional affixes is a prototypical trait of creole languages. In most (if not all) Atlantic creoles all verbal inflections disappeared (Ansaldo, Matthews & Lim 2007; see (1) and (2)):

(1) He walk-s (English)

(2) He walk (Hawaiian Creole)

This can be allotted to the properties of '-s' in that context. Namely, in the English sentence the verbal inflection '-s' does not significantly contribute to the overall meaning of the sentence as the same properties are already expressed by the personal pronoun 'he'. Hence, it is assumed that the verbal inflection has little semantic content and is therefore less likely to get selected in a situation of contact. Although other properties of the verbal inflection, (e.g., its perceptual salience) are likely to contribute to its loss as well, it is improbable that the morpheme would have disappeared if it were functionally more salient (Ansaldo 2009).

In the literature on language contact, salience is often reported as an explaining factor of change (Trudgill 1986, Hinskens 1996, Deumert 2003, Woolard 2008). However, the salience of a linguistic feature is defined in a myriad of ways and often partly or wholly overlaps with either its type or token frequency (Grondelaers and Geeraerts 2003), its perceptual prominence (Mufwene 1991) or its ease of activation (Schmid 2007). This shows a need to be explicit about exactly what is meant by the term 'salience'. Although there is a natural appeal to the explanation that salient linguistic features are more likely to prevail in language contact, without a proper definition of salience

this hypothesis gives rise to circularity. That is, one needs to address why certain features become salient and therefore more prone to linguistic change. Also, was it salience that caused a feature to get selected by a speaker or has the feature become salient because it got selected?

In this paper, the focus is on functional or semantic salience (coined as semantic load), which is defined as: 'the contribution of a single word or morpheme to the overall sentence meaning'. It differs from the definitions of salience previously mentioned in that it is context specific (a lexical item can be of importance in one sentence but not in the other), it only applies to the semantic function of words and it does not incorporate perceptual salience. The notion of semantic load is coined based on the analogy of functional load (Gilliéron 1918) defined as the degree to which certain linguistic features are important in making distinctions in a language. Martinet (1952) elaborated on this notion claiming that sound changes are less likely to occur with phonemes that have high functional load. Semantic load is quantified using a neural network model that incorporates principles from distributional and compositional semantics. It is hypothesized that in situations of competition between features, features with high semantic load are more likely to get selected than features with low semantic load. In other words, is semantic load a predictor of feature selection and accordingly of the direction of language change?

The outline of this paper is as follows. In the following section, it is explained how a neural network model can be employed to learn representations of words and sentences and how these representations can be manipulated to get a measure of a word's contribution to the sentence meaning. Section 3 outlines the results from the model. Finally, this paper concludes with several case examples of how the quantified notion of semantic load can explain phenomena observed in cross-linguistic influence and contact languages and discusses the findings and directions for future research.

2. Quantifying semantic load

The degree of semantic load is hypothesized to be one of the cognitive constraints that might influence the selection of certain features (i.e., variants) over others in the mind of an individual speaker. However, the notion of semantic load is hard to quantify in a principled manner. Especially in the case of competition between features from different linguistic systems, it is hard to quantify semantic content in a manner that allows for cross-linguistic comparisons. In this section, a bilingual computational model is introduced that uses concepts from the field of lexical distributional semantic and

compositional semantics to arrive at a notion of semantic load that is quantifiable and can be applied cross-linguistically. The working definition of semantic load is as follows: 'the contribution of a single word or morpheme to the overall sentence meaning'. First, in section 2.1 it is explained how distributional representations are used in computational models to learn to represent words in a meaningful way. Then, in section 2.2, the principle of compositionality is introduced and it is shown how it can be applied computationally to get representations of larger phrases and sentences. Lastly, in section 2.3, the architecture of the bilingual neural network is outlined and it is shown how the contribution of a word to the sentence meaning can be quantified.

2.1 Distributional representations of meaning

Computer models work with symbols, and accordingly, depend upon symbolic representations. In computational linguistics and Natural Language Processing (NLP) one of the major challenges is to teach computers to understand human languages. For this to function, computers must learn representations of meaning that are presumably similar to the way humans represent meaning. A very successful approach in both computational linguistics and NLP has been that of using distributed representations:

"The terms distributional, context-theoretic, corpus-based or statistical can all be used (almost interchangeably) to qualify a rich family of approaches to semantics that share a "usage-based" perspective on meaning, and assume that the statistical distribution of words in context plays a key role in characterizing their semantic behaviour." (Lenci, 2008)

The Distributional Hypothesis (DH) posits that at least certain aspects of the meaning of lexical items are dependent on the distribution of those items (Lenci, 2008). That is, the meaning of a word is in part (or as argued by some: completely) determined by the surrounding context (Harris 1951, Firth 1957). Hence, word meaning can be represented as the co-occurrence counts in text corpora because it is assumed that words that occur in similar contexts have similar meanings. For example, the word 'cup' and 'mug' will occur in similar contexts and will therefore have similar co-occurrence counts and hence, similar distributed representations. Within corpus and usage-based linguistics the DH has played a significant role, owing to an elegant match with other central notions in these frameworks, such as context sensitivity and statistical patterns of usage. However, opponents of the distributional hypothesis have argued that meaning representations must in some way be anchored to extra-

linguistic entities, either to objects in the world or embodied conceptual representations.

Computational models make extensive use of the statistical properties of language and accordingly, the use of distributed representations to represent meaning provides an easy-to-operationalize solution to the longstanding puzzle of how to formalize meaning (Lenci, 2008). Computational models that employ distributed representations in most cases generate those representations from large text corpora using word co-occurrence statistics. In the image below an example is given of how frequency data from a corpus is used to get a numeric representation of the word **lamb**.

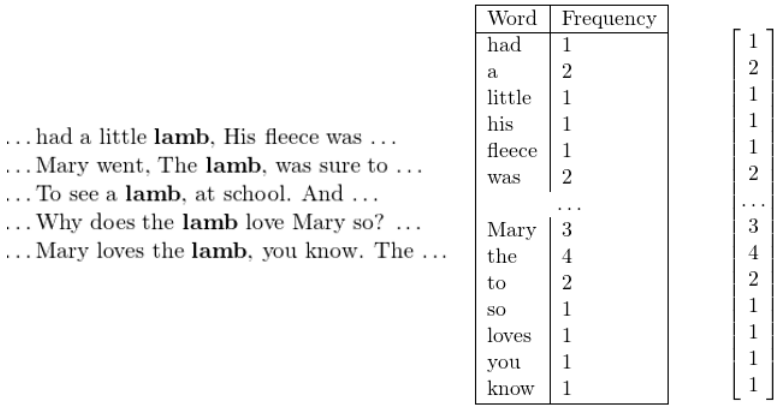

Figure 1 Co-occurrence vector of the word 'lamb' using a window size of 3

As can be gleaned from the Figure above, the window size (3, in this case) reflects how many words before and after the target word are counted as context of the target word **lamb**. A simple matrix that lists all words present in the corpus and their co-occurrence counts will result in vector representations of these words. However, this method will produce vector representations of very large dimensions which is why most models of vector representations use dimensionality reduction techniques (see e.g., GloVe (Pennington et al. 2014)).

Models that use distributed vector representations have been successful in detecting semantic similarity (Turney 2006), making predictions about semantic priming (Lund et al. 1995), retrieving information from large texts (Manning et al. 2008) and disambiguating word-senses (Padó and Lapata 2007). By depicting word vectors as points in high dimensional semantic space, the position of a word is determined by its distribution in different contexts and the distance between words in vector space corresponds to their semantic distance (Lenci 2008). Cognitive scientists have argued that there is

psychological and neurological evidence for the cognitive plausibility of distributional representations. For example, Mitchell et al. (2008) present a computational model that predicts the fMRI activation patterns when participants are thinking about concrete nouns by using distributed representation of those nouns. During training, the model extracts semantic features from word co-occurrence in a large text corpus and subsequently predicts the neuronal activation patterns by weighting the contributions of each of the semantic features. Hence, the model assumes "a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings" (pp. 1194). Individual models were trained for each of the participants and all models could predict – with an accuracy significantly above chance level – neuronal activation patterns for newly presented nouns.

Turvey and Pantel (2006) argue that vector space models are the most successful (computational) approach to semantics to this point. Properties of distributed vector representations make them attractive for modelling language and meaning, namely they are context-sensitive, gradual, dynamic and not determined by previously defined features. The success of computational models that use distributed representations suggest them to be a promising tool for further research.

2.2 Compositional semantics

The meaning of a sentence is not simply the sum of the meaning of its parts. Constructions that are composed of the same parts, can have different meanings (see (1) and (2)).

1) The book with the cover.

2) The cover with the book.

Partee (1984), to account for this aspect of language, adheres to the Principle of Compositionality: "the meaning of a whole is a function of the meanings of the part and of the way they are syntactically combined" (pp. 153). What logically follows from this principle, it that meaning cannot be constructed without the use of syntax. Montague (1970), the founder of 'formal semantics', argued that 'ordinary' language can be described in formal terms and just as precisely as any formal language. It is generally accepted that the meaning of a sentence is dependent on its structure and evidence for the compositionality of language comes from both psycho- and neurolinguistic research. Recent studies on processing complexity revealed that sentences that have a higher

syntactic complexity (i.e., they consist of multiple decomposable layers) incur more brain activation compared to sentences of similar length that do not have this multi-layered compositional structure (Santi and Grodzinsky, 2010).

Computational models have used the principle of compositionality to learn to represent meaning of phrases and sentences (Socher et al., 2013; Hermann and Blunsom, 2014; Le and Zuidema, 2015). As these models work with symbolic languages, the formal aspect of Montague's grammar provided clearly defined tools to construct meaning. The most straightforward method to arrive at a representation of a sequence of words, is by simply adding the vector representations. The compositional vector is derived by adding up the multiple component vectors (Landauer and Dumais, 1997). Another method is that of component wise multiplication, in this case the component features of the vectors are multiplied. This results in a boost for features that are shared among the two words, while those features that are not shared disappear from the compositional representation. However, these methods do not take into account order, and hence, 'Horses run' and 'Run horses' would get the same representation.

Models that rely more heavily on the formal semantics framework are better able to account for compositional structure. Baroni and Zamparelli (2010) propose a model in which the type of syntactic relation between words determines the type of operation that is performed on them to arrive at a representation of their composition. Based on insights from formal semantics, Baroni and Zamparelli (2010) argue that adjectives as in the adjective-noun pair 'old dog' can be treated as functions that operate on nouns. They use matrix - vector multiplication where adjectives are represented as matrices (i.e., an ordered set of weights) and are applied to the vector representation of the noun. The outcome of that operation results in a compositional vector representation. Socher et al. (2012) extend this idea to entire sentences by assigning a matrix and vector representation to every word in a sentence. In their matrix-vector recursive neural network, the syntactic structure of the sentence determines how words are combined to form larger phrases. This model manages to capture very fine-grained meanings of compositional structures and performs well in tasks such as sentiment analysis and text categorization. The model employed to quantify semantic load – introduced in detail in the following section – uses a similar approach to Socher et al. (2012) to learn to represent the meaning of phrases and sentences.

2.3 A bilingual Recursive Neural Network for sentence alignment

2.3.1 Architecture

The model used to quantify semantic load is an adaptation of Hermann and Blunsom's (2014) multilingual model for compositional semantics. In this model, sentence representations are learned by aligning the representations of parallel sentences in a “shared multilingual semantic space” (pp. 1) while at the same time dissimilar sentences are forced to have a certain distance between them. One of the advantages of this approach is that no external information is required to semantically ground the sentences (Hermann and Blunsom, 2014). It is the parallel sentence in the other language that is used as its grounding. In Figure 2, a schematic representation of the architecture is depicted, where f(a) (lower box with red dots) represents a sentence in Language A (e.g., Mandarin) and g(b) (upper box with red dots) is its parallel translation in Language B (e.g., English).

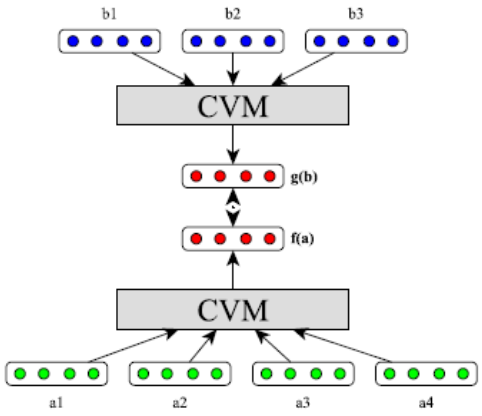

Figure 2 Model with parallel input sentences a and b. The model minimises the distance between the sentence level encoding of the two sentences (from Hermann and Blunsom 2014)

The box in Figure 2 that is marked `CVM' can represent any kind of compositional vector model to generate sentence level representations. Hence, the neural network model works as follows: (1) first, individual sentence representations for the two parallel sentences are computed using any type of compositional vector model (henceforth: CVM) in the respective languages, (2) the two sentence representations in language A and language B are compared and the error (how distant they are in ``shared multilingual semantic space'') is computed, (3) the connection weights of the CVMs in both languages are optimized to reduce the error in the next trial. Additionally, some non-parallel sentences are presented to the model and it is ‘taught’ to keep a large margin of distance between these representations. This type of architecture follows a supervised paradigm as it receives feedback after each

sample of parallel sentences, specifying whether the set of sentences is supposed to have similar or dissimilar representations in joint semantic space.

The CVM model used to learn compositional representations of sentences in both Mandarin and English is the Inside-Outside Recursive Neural Network (IORNN; Le and Zuidema 2014a). The IORNN builds up phrase and sentence representations in a bottom-up fashion (as in Socher et al. 2012), but also in a top-down fashion. For many linguistic tasks, we need information that can flow in both directions (Le and Zuidema, 2014a), as the meaning of a word is determined both by contextual information and the actual lexical item. The advantage of the IORNN is that both component (inside) and context (outside) representations are able to interact. The compositional structure of the sentences is determined by its dependency tree (see Figure 3; Chen and Manning 2014). Dependency trees are semantically based structures that represent the relationship between words in a sentence. Models of compositional semantics have shown to benefit from dependency structures (see e.g., Nivre et al., 2008) as compared to the more traditional binary constituent tree structures.

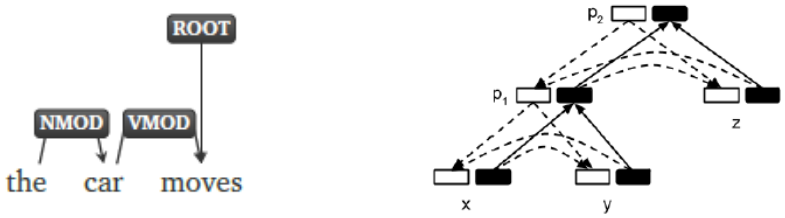

Figure 3 Dependency tree of the phrase 'the car moves' (left) and a schematic representation of the inner (black) and outer (white) representations in the IORNN (x = 'the'; y = 'car' and z = 'moves') (right)

The IORNN learns a vector representation of the sentence 'the car moves' by first computing a compositional inner and outer representation of 'the car' (see Figure 4 below). The inner representation is defined compositionally by the parts it comprises (e.g., 'the' and 'car' for the representation of 'the car') and the outer representation is defined compositionally by what it is surrounded with (e.g., the outer representation of 'the car moves' and the inner representation of 'moves'). Note that different weights are assigned to the head and the modifier and to the parent and the sister node. As the highest node in the dependency tree (i.e., the root) does not have a surrounding context to compute an outer representation, the outer representation of that node is assigned a random value that is adjusted during training (Le and Zuidema, 2014b). To arrive at a full representation of the phrase, the inside representation of the sentence 'the car moves' is computed by the weighted sum of the vector representation of 'the car' and 'moves'.

$$i_{\text{the car}} = f(W_{\text{head}} \cdot i_{\text{car}} + W_{\text{nmod}} \cdot i_{\text{the}} + b)$$
$$o_{\text{the car}} = g(W_{\text{parent}} \cdot o_{\text{the car moves}} + W_{\text{sister}} \cdot i_{\text{moves}} + b)$$
$$i_{\text{the car moves}} = f(W_{\text{head}} \cdot i_{\text{moves}} + W_{\text{vmod}} \cdot i_{\text{the car}} + b)$$

Figure 4 Computation of inner and outer representation of the NP 'the car' and inner representation of 'the car moves'

Following Le and Zuidema (2014b), the model is trained to predict missing words in a sentence by using the representation of the surrounding context (i.e., the outside representations). In Le and Zuidema (2014b), it was shown "that the IORNN is capable of judging very well how different words fit the context created by a surrounding utterance" (p. 6). Hence, the multilingual IORNN model for compositional semantics is trained on the dual task of learning to align parallel sentences in Mandarin and English and predicting missing words based on its surrounding context.

2.3.2 Training and obtaining the semantic load

The multilingual sentence alignment model (Hermann and Blunsom, 2014) is trained on the UM-corpus (Tian et al., 2014), which contains over two million parallel Mandarin Chinese-English sentences in the domains of News, Education, Law, Thesis, Subtitles, Spoken, Microblog (Twitter data) and Science. Data from all domains is used as input for the model. The corpus data was split into a training and testing part for possible further evaluation of the results. The dependency trees were obtained using the dependency parser of Levy and Manning (2003) for Mandarin and Chen and Manning (2014) for English. All model weights and word vector representations were randomly initialised and are learned during the training phase. The dimensionality of the word representations was set to d=100 (i.e., all vectors have a length of 100), as in an earlier model no major differences were found between word representations of d=100 and d=200. For each aligned sentence pair, 10 randomly selected non-aligned sentence pairs were used as noise samples.

As explained in the previous section, the IORNN follows a dual task training paradigm. The model is trained on both missing word prediction and aligning the Mandarin and English sentence representations in semantic space. The respective tasks are weighted 0.4 and 0.6 respectively. Thus, the task of aligning the parallel sentences is regarded as slightly more important than missing word prediction. In this case, the model tries to minimize both the error for all semantically aligned sentences (weighted 0.6) and the error of word prediction (weighted 0.4).

After training the multilingual model on the dual task of sentence alignment and missing word prediction it is possible to extract vector representations for all words and sentences in the model. With these representations, it is possible to quantify the contribution of individual words to the sentence meaning. The semantic load is computed by calculating the distance (using cosine similarity) between the vector representation of the sentence with and without that word. As is shown in Figure 5, to calculate how much 'the' contributes to the sentence meaning, the vector representation of 'the' is replaced by an empty vector. This empty vector contains only zero values and has arguably lost all semantic information it carried. Cosine similarity is a commonly used measure to estimate the proximity of two vectors in high-dimensional space and is used to compare the sentence representations with and without the empty vector.

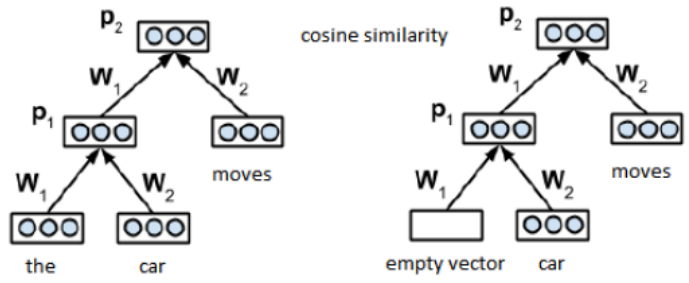

Figure 5 The semantic load of 'the' is calculated by replacing it with an empty vector and measuring the distance between the sentence representation with and without 'the'.

3. Results

In Figure 6 and 7 below, results for two parallel sentences are depicted with their dependency structure. As can be gleaned from Figure 6, 'similar' in this sentence gets a semantic load of 0,22, 'rules' receives a semantic load of 0,11, 'have' a semantic load of 0,09, and so forth.

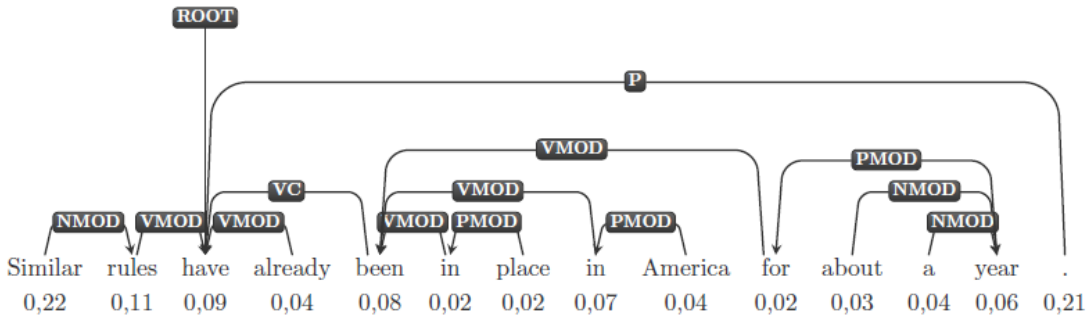

Figure 6 Semantic load per word for an example sentence in English

The dependency structure might seem daunting at first, but it simply shows the compositional structure of the sentence. That is, it shows which words are dependent on which words and how they are ultimately related to the root of the sentence. For example, in Figure 6, 'America' is dependent on 'in', which is dependent on 'been', which in turn is dependent on the root of the sentence 'have'.

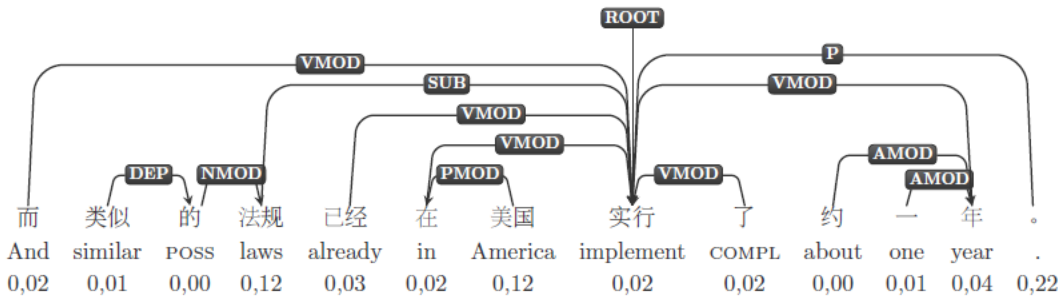

Figure 7 Semantic load per word for an example sentence in Mandarin Chinese

As can be gleaned from Figure 7, the dependency structure and the distribution of semantic load differs in the Chinese sentence. The root in this sentence has many direct dependents and there are fewer words deeply embedded in the structure.

3.1 Aggregated results

The results for individual sentences are informative when comparing the load of individual words in context. However, aggregated results can be helpful when wanting to compare word categories and general patterns. In Table 1, the average load for all main word categories are outlined with on the left the categories for Mandarin and on the right the categories for English. The part-of-speech tags are assigned by the Stanford parser when obtaining the dependency trees for the sentences in the corpus (Chen and Manning 2014), but are not presented as input to the model. That is, the neural network model does not have this type of information to learn to represent the meaning of words and phrases.

Part of speech (Mandarin)	Average load	Part of speech (English)	Average load

Pronoun	0,14	Modal verb	0,13
Punctuation	0,13	Personal pronoun	0,13
Subordinate clause marker	0,12	Verb (3SG present)	0,11
Interjection	0,11	Punctuation	0,10
Verb copula	0,09	Verb (past tense)	0,10
Sentence final particle	0,08	Coordinating conjunction	0,10
Adverb	0,07	Verb (non-3SG present)	0,09
Verb	0,06	Adverb	0,07
Ordinal number	0,06	Possessive personal pronoun	0,07
Proper noun	0,05	Interjection	0,06
Preposition	0,05	Verb (past participle)	0,06
Aspect particle	0,04	Proper noun	0,05
Predicative adjective	0,04	Preposition/subordinating conjunction	0,05
Demonstrative	0,04		0,05
Noun	0,04	Noun	0,05
Genitive DE	0,04	Verb (gerund)	0,05
Adjective	0,03	Verb (root form)	0,05
Measure word	0,03	Cardinal number	0,05
Clause conjoiner	0,03	Determiner	0,04
		Adjective	0,03
		Possessive marker	

*Table 1 Main word categories with their average semantic load in Mandarin (left) and English (right***)**

The categories in Table 1 differ noticeably for the two respective languages but some common patterns can be found. For example, in both Mandarin and English pronouns and punctuation get a high average semantic load. A perhaps surprising result is the relatively low average semantic load of nouns, as these are intuitively important for the overall sentence meaning.

Besides averaging the results per word category, another possibility is to average the contribution of a specific word to the sentence meaning in all contexts. Table 2 depicts the average semantic load of the words ‘but’, ‘or’ and ‘and’ in both Mandarin and English. As clearly demonstrated, in both languages ‘but’ receives the highest semantic load.

Word (Mandarin)	Average load	Word (English)	Average load
但 (but)	0,19	but	0,27
或 / 或者 (or)	0,06 / 0,05	or	0,08
和 (and)	0,02	and	0,06

Table 2 Average semantic load in Mandarin and English for the words 'but', 'or' and 'and'

In the following graph, the relationship between depth of embeddedness and semantic load is depicted. Depth of embeddedness refers to the hierarchical distance of a word to the root of the sentence (assigned by the dependency parser). For example, in the English sentence in Figure 6, 'America' has a depth of 4 because it is three steps away from the root (and the root has a depth of 1). Figure 8 shows that there is a relationship between the depth of a word and the semantic load it gets in the model. Words that are less embedded in the compositional structure get a higher semantic load than words that are more deeply embedded. Moreover, there appears to be a structural difference between English and Mandarin with respect to embeddedness. As already noted when discussing Figure 7, Mandarin seems to have fewer sentences with deeply embedded words (with a maximum depth of 15 compared to 22 in English) than the English data.

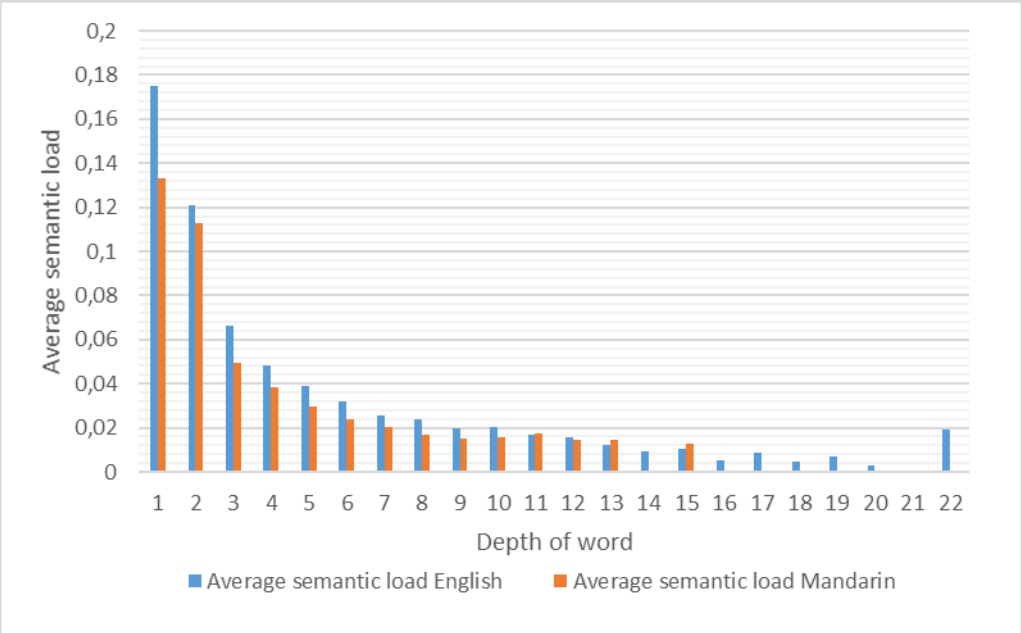

Figure 8 Depth of word and average semantic load for English (blue) and Mandarin (orange)

4. Discussion and further research

In this paper, a computational method has been introduced to quantify the contribution of a single word to the overall sentence meaning. Semantic load has been described both conceptually and methodologically, and is proposed as a predictor of language change. That is, within the framework of competition and selection in the feature pool (Mufwene, 2001) it is hypothesized that words with a higher semantic load are more likely to get selected than words with a low semantic load. By getting selected, these words are more prone to change as they acquire new functions or adapt to the new environments in which they are used.

As described in the introduction, the borrowing hierarchies presented in Matras (2009) provide an excellent testing ground for the semantic load hypothesis. The results presented in the previous section show that 'but' receives a high semantic load in both English and Mandarin while 'or' and 'and' follow far behind. Hence, in a situation of contact 'but' would make a strong competitor and would more likely get selected than 'or' and 'and', following the hierarchy 'but > or > and' presented in Matras (2007). Matras accounts for this cross-linguistic pattern of borrowing by referring to the property of 'but' as belonging to the presupposition domain. That is, 'but' is contrasting a shared presupposition, a feature (i.e., that of expressing contrast) which is argued to be highly susceptible to borrowing. This intrinsic semantic-pragmatic feature seems to have been picked up by the computational model and is reflected in its semantic load.

The model is trained to align parallel sentences in Mandarin and English and can therefore make predictions about language change in societies where these languages come into contact. Mandarin and English have contributed to the mixed variety of Colloquial Singapore English and it is hypothesized that the semantic load as measured by the model can predict the selection of certain features in this variety. One of the distinctive features of Singapore English is the widespread use of sentence final particles (Wong, 2004). Table 1 in the previous section shows results from the English-Chinese bilingual model in which sentence final particles get assigned a relatively high load. Accordingly, sentence final particles would make up strong competitors and would therefore be more likely to emerge in individual contact varieties, such as Colloquial Singapore English. Pragmatic particles are 'notorious' for emerging in contact varieties (Siemund and Kintana 2008) which is usually

accounted for by their property of belonging to the discourse domain. Similar to 'but', the intrinsic semantic-pragmatic function of sentence-final particles might have contributed to their high semantic load. Alternatively, their high semantic load can be explained by the embeddedness of these particles in the sentence structure. In the model, words that occur as roots or direct dependents of the root receive a high semantic load. As sentence final particles have scope over the whole phrase, they are always directly dependent on the root. Hence, it might be the case that their high load depends partially on this property, which in turn might explain their frequent appearance in contact varieties.

Woolard (2008), in a paper on sociolinguistic salience and language change points to frequent recurrence of personal pronouns as markers of social differentiation (e.g., the phonological change of 'I' to 'ah' as a Texan identity marker). She suggests that linguistic changes tend to occur more often within lexical items (such as personal pronouns) that are foregrounded in interaction. Errington (1985) argues that personal pronouns are more likely to become pragmatically salient because they are referential, they refer to persons and are indexical. In studying changes in the Javanese politeness, Errington finds that changes have occurred more rapidly in personal pronouns than in other lexical items. In the model, both Mandarin and English personal pronouns receive a high semantic load. This seems intuitive because 'I want to do it' compared to 'You want to do it' are very different in meaning. Also, in the task of aligning two parallel sentences semantically, the model might easily single out pronouns as they form a small closed class. What the results from this model seem to suggest is that personal pronouns are also semantically salient, which can further account for their recurrent role in linguistic change as discussed in Woolard (2008) or in the bilingual idiolectal varieties discussed in Sorace and Filiaci (2006).

As reported in the previous section, nouns got assigned a relatively low semantic load in both Mandarin and English. A general pattern that emerges from the analysis, is that closed class words with a clear semantic function get on average a higher semantic load than open class words such as nouns and verbs. This can be explained by the way in which the model learns word representations. During training, the model needs recurrent input of a word to learn an accurate representations. Many nouns in the corpus might only recur several times and their representations are therefore not optimally learned. Because of the high number of low frequency nouns in the corpus, the average semantic load of nouns is also relatively low.

An important point that possibly needs more explanation is the somewhat paradoxical relationship between having high semantic load and being more prone to language change. At first intuition, it might be more plausible to assume that words with a high semantic load are more stable in a given language and hence, less prone for change. However, an important aspect to bear in mind is that change occurs due to contact with different idiolects and having a high semantic load implies being a strong competitor. Weak competitors will disappear in the process of replication while strong competitors get adopted by new speakers and will acquire new functions.

The results presented in this and the previous sections are just an illustration of what can be done with the vast amount of data that the computational model gives as output. With the vector representations of words and phrases, it is possible to further analyse the organization of words in semantic space and get a better understanding of how it deals with bilingual representations. For example, would the model learn to represent the word for 'I' similarly to its Mandarin counterpart '我'? Also, in the case examples above, the quantified notion of semantic load reflects the average degree to which a word contributes to the meaning of the overall sentence. However, a word in one context can contribute strongly to the overall meaning of the sentence while in another context the same word might contribute just marginally to the overall meaning. This contextual variation disappears when averaging over all contexts. In online competition and selection in the mind of a speaker, it might be the case that in high semantic load contexts the word is a strong competitor and hence, gets selected while in low semantic load contexts another competitor 'wins'. With the data presented above, it is possible to extract the semantic load of a word per context. Moreover, one can present the model with any type of language combination, provided that there is a parallel corpus available and a dependency parser to determine the dependency structure of the sentences in the corpus. Future research should also look into the applicability of the notion of semantic load to other areas in linguistics. Possible directions are psycholinguistic processing studies in which it is tested whether elements with high semantic load are harder to process or more likely to get selected when performing a task with high working memory demands. Studies on language acquisition could possibly look at the order of acquisition of words with high and low semantic load.

5. References

Aboh, E. O. (2015). The Emergence of hybrid grammars: Language contact and change. Cambridge University Press.

Aboh, E. O., & Ansaldo, U. (2006). The role of typology in language creation. *Deconstructing creole*, 39-66.

Ansaldo, U. (2009). Contact language formation in evolutionary terms. In: Aboh, E., Smith, N. (Eds.), *Complex Processes in New Languages*. John Benjamins, Amsterdam/Philadelphia, 265–289

Ansaldo, U., Matthews, S., & Lim, L. (2007). *Deconstructing creole* (Vol. 73). John Benjamins Publishing. Amsterdam.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193. Association for Computational Linguistics.

Chen, D., & Manning, C. D. (2014, October). A Fast and Accurate Dependency Parser using Neural Networks. In *EMNLP* (pp. 740-750).

Croft, W. (2000). Explaining language change: an evolutionary approach. Pearson Education.

Deumert, A. (2003). Markedness and salience in language contact and second language acquisition: evidence from a non-canonical contact language. *Language Sciences*, 25(6),561–613.

Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a complex adaptive system* (Vol. 3). John Wiley & Sons.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635-653.

Errington, J. J. (1985). On the nature of the sociolinguistic sign: Describing the Javanese speech levels. *Semiotic mediation: sociocultural and psychological perspectives*, 287-310.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955.

Grondelaers, S., & Geeraerts, D. (2003). Towards a pragmatic model of cognitive onomasiology. *Cognitive approaches to lexical semantics*, *67*, 92.

Gilliéron, J. (1918). Généalogie des mots qui d´esignent l'abeille d'apres l'atlas linguistique de la france. Paris: É. Champion.

Harris, Z. S. (1951). Methods in structural linguistics.

Hinskens, F. (1996). Dialect levelling in Limburg: Structural and sociolinguistic aspects (Vol. 356). Walter de Gruyter.

Kirby, S. (1998). Fitness and the selective adaptation of language. In: Hurford, J. R., S.-K. M. and C., K. (eds.), *Approaches to the Evolution of*

Language: Social and Cognitive Bases. Cambridge University Press, Cambridge.

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

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Le, P., and Zuidema, W. (2014a). The inside-outside recursive neural network model for dependency parsing. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 729-739.

Le, P., and Zuidema, W. (2014b). Inside-outside semantics: A framework for neural models of semantic composition. In *NIPS 2014 Workshop on Deep Learning and Representation Learning*.

Le, P. and Zuidema, W. (2015). Compositional distributional semantics with long short term memory. In *Joint Conference on Lexical and Computational Semantics*, 10–19.

Levy, R. and Manning, C. D. (2003). Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 439–446. Association for Computational Linguistics.

Lund, K., C. Burgess, and R. A. Atchley. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society*, 660–5. Cognitive Science Society, University of Pittsburgh.

Manning, C. D., P. Raghavan, and H. Schütze. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Martinet, A. (1952). Function, structure, and sound change. *Word*, 8, 1-32.

Matras, Y. (2007). The borrowability of structural categories. *Empirical Approaches to Language Typology*, *38*, 31-73.

Matras, Y. (2009). *Language contact*. Cambridge University Press.

Matras, Y. (2011). Explaining convergence and the formation of linguistic areas. *Geographical Typology and Linguistic Areas*, 143-160.

McWhorter, J. (2001). The world's simplest grammars are creole grammars. *Linguistic typology*, *5*(2/3), 125-166.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320* (5880), 1191-1195.

Montague, R. (1970). Universal grammar. *Theoria*, 36(3):373–398.

Mufwene, S. S. (1991). Pidgins, creoles, typology, and markedness. *Development and structures of creole languages*, 123-143.

Mufwene, S. S. (2001). *The ecology of language evolution*. Cambridge University Press.

Mufwene, S. S. (2008). Language evolution: Contact, competition and change. Bloomsbury Publishing.

Nivre, J., Hall, J., and Nilsson, J. (2008). Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, pages 49–65.

Padó, S., and M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2). 161–99.

Partee, B. (1984). Compositionality. *Varieties of formal semantics*, 3:281–311.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, 14, 1532–1543.

Roeper, T. (1999). Universal bilingualism. *Bilingualism: language and cognition*, 2(03):169–186.

Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. *The Oxford handbook of cognitive linguistics*, 117-138.

Siemund, P., & Kintana, N. (Eds.). (2008). *Language contact and contact languages* (Vol. 7). John Benjamins Publishing.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. ACL, 1201-1211

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013). Parsing with compositional vector grammars. In *ACL (1)*, 455–465.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research,* 22:3, 339-368.

Sorace, A., & Serratrice, L. (2009). Internal and external interfaces in bilingual language development: Beyond structural overlap. *International Journal of Bilingualism*, 13:2, 195-210.

Steels, L. (2000). Language as a complex adaptive system. *Parallel Problem Solving from Nature PPSN VI* (pp. 17-26). Springer, Berlin Heidelberg.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., and Oliveira, F. (2014). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In: *LREC,* 1837-1842.

Trudgill, P. (1986). *Dialects in contact*. Blackwell.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics* 32(3). 379–416.

Wong, J. (2004). The particles of Singapore English: A semantic and cultural interpretation. *Journal of Pragmatics*, *36*(4), 739-793.

Woolard, K. A. (2008). Why dat now?: Linguistic-anthropological contributions to the explanation of sociolinguistic icons and change. *Journal of Sociolinguistics,* 12(4):432–452.