The Rationality of Semantic Change

Omer Korat
Stanford University
omerk@stanford.edu

Abstract

This study investigates the mutual effects over time of semantically related function words on each other's distribution over syntactic environments. Words that can have the same meaning are observed to have opposite trends of change in frequency across different syntactic structures which correspond to the shared meaning. This phenomenon is demonstrated to have a rational basis: it increases communicative efficiency by prioritizing words differently in the environments on which they compete.

1 Introduction

In this paper I propose that as words immigrate to new syntactic environments over time, they tend to push out words that populated these environments prior to immigration. This results from a process of reasoning over the lexicon, in which speakers choose among lexical alternatives in a way that optimizes communicative utility. In particular, I focus on discourse markers (DMs) and prepositions.

Computational modeling of historical change has seen increased popularity in recent years (Xu and Kemp (2015); Kulkarni et al. (2015); Hamilton et al. (2016b,a) (distributional semantics), Basile et al. (2016) (n-gram models), Schaden (2012); Deo (2015); Yanovich; Enke et al. (2016); Ahern and Clark (2017) (evolutionary game theory)). In this paper, parsed historical corpus evidence is used to quantify existing claims about semantic change, some of which have not been empirically assessed (Bréal, 1897; Ullman, 1962; Traugott and Waterhouse, 1969; Clark and Clark, 1979; Sweetser, 1991; Traugott, 1995; Traugott and Dasher, 2002). Data were collected from the Penn Parsed Corpora of Early Modern English (Kroch and Delfs, 2004) and the Parsed Corpora of Early English Correspondence (Taylor and Nevalainen, 2006).

Specifically, I conduct a quantitative investigation of the manifestation of the principle of contrast (Paul, 1898; Bréal, 1897; Clark, 1990; De Saussure, 1916) in semantic change with abstract terms. According to the principle of contrast, difference in form between two lexical items (phonology/orthography) leads to difference in semantics (Section 2.1). The effects of this principle on semantic change have been studied in the literature (Xu and Kemp, 2015; Hamilton et al., 2016a,b), but previous studies have focused on content words using unambiguous bag-of-words based word vectors, while this study focuses on function , employing syntax-based representations which take into account the structural position of the word. These representations allow for a word to have multiple uses, unlike in the bag-of-words approach. Such representations are better suited to study the distribution of function words, because the meaning of these words can vary based on its syntactic position. For example, *so*, when it appears as a complementizer, is a discourse marker, but when it appears inside a verb phrase is a manner adverb. Such distinctions can only be captured with ambiguous representations that take into account structural information beyond the bag-of-word level.

I argue (Section 2.2) that (a version of) this principle leads to the prediction that when some word immigrates to a new environment, it will compete with other words in that environment, leading to older alternatives becoming less frequent in it. This prediction is tested and verified in Section 4.

The environments considered in this paper are defined syntactically, utilizing manually parsed corpora. Semantic change is measured using hand-crafted distributional syntactic features. The advantage of using syntactic features include (i) the ability to distinguish between different uses of lexically ambiguous words and (ii) ability to

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 151–160
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

utilize syntactic information which is absent from unannotated corpora.

This approach can supplement other quantitative approaches, such as distributed semantics, in which all the uses of a word are compressed into one measure (its point in vector space). In previous studies, word vectors were built using pure bags-of-words, and therefore were unable to distinguish different uses of the same word that differ in their syntactic position. Thus, they were not measuring the relative changes in the frequencies of each different use over time. This limitation is addressed in this study by the use of syntactic features taken from parsed corpora.[1]

Hand crafted syntactic features are therefore well suited to analyzing the development of discourse markers. DMs tend to have many different uses, each with its own distributional properties, and capturing them requires tapping into meta-linguistic information which does not exist in unannotated corpora.

2 The Principle of Contrast

2.1 Background

The principle of contrast states that any two forms contrast in meaning (Bréal, 1897; Paul, 1898; Clark, 1990; De Saussure, 1916). The principle is based on the intuition that when speakers choose linguistic expressions, they do so because they mean something they would not mean by some other expression. Part of the meaning of expressions emerges due to the contrast with alternatives. If a given meaning M is already associated with a well-established form F_1, when the speaker uses a different form F_2, the addressee infers that the speaker did not mean M, since it is common knowledge that the speaker assumes that the addressee can readily compute a unique meaning for F_2.

Note that difference in form (phonology or orthography) can motivate difference in distribution but not in meaning. As Wasow (2015) points out, despite the appeal of the idea of contrast (as formulated by Grice (1975)), language is, in fact, ambiguous. This is evident for example by the fact that many sentences have multiple possible parses, and that speakers sometimes ask for clarification about the sense in which a word was used. Moreover, see the discussion in Section 4.1 for evidence that challenges the principle of contrast.

Contrast should be seen as one motivating force among many in semantic change. Wasow discusses other possible factors which might motivate ambiguity (e.g. such as reliance on speakers' reasoning faculties in order to save time) which lead speakers to leave out information, resulting in ambiguity. Such variables might operate in parallel to the principle of contrast, leading to ambiguity arising in some situations but not others. That is, contrast in the phonology of two words motivates speakers to use those words differently. This can entail difference in semantics, as is claimed by e.g. Clark and Gathercole, but it can also entail distributional difference, that is, difference in likelihood to appear in certain environments, based on social, syntactic or pragmatic conditions.

2.2 Contrast and Function Words

Here, we adopt a relaxed version of the principle of contrast according to which, word pairs that can mean the same thing in some environment will tend to have different distributions in that environment. This version can be thought of as a natural consequence of general logical principles and Bayesian inference, as in (Hobbs, 1985; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; De Jaegher and van Rooij, 2014; Ahern and Clark, 2017). Ahern and Clark (2017) demonstrate how rationality principles can account for semantic drift phenomena. This shows that as in the principle of contrast, the distribution of words is modulated at least partially by rational communicative heuristics aimed to save cognitive effort. Speakers exploit these facts to select which words to use when, thus increasing communicative efficiency and saving cognitive effort. In Rational Speech Acts Theory (RSA, Frank and Goodman (2012)) these notions have been generalized to account for pragmatic inference in the general case by assuming that the probability of an utterance is proportional to its information gain over its cost.

This applies directly to the case at hand: by prioritizing function words differently in different environments, speakers can increase the information gain of their utterances, thus reducing the expected cost of communication. Formally, let u_1 and u_2 be identical and synonymous utterances that differ in

[1] Note that the same result could be accomplished using syntax-based word vectors (Padó and Lapata, 2007; Weir et al., 2016; Antoniak and Mimno, 2018) or corpus-based semantic models (Baroni and Lenci, 2010; Petrolito and Bond, 2014). However, such corpora typically do not include historical data, unlike the corpora used in this study.

one word. Let this word be w_1 in u_1 and w_2 in u_2. Now, assume u_1 and u_2 compete on some environment E (syntactic, pragmatic, social, etc'). If speakers do not make a distinction between the uses of w_1 and w_2 in E, then there is no reason to store both the E-use of w_1 and the E-use of w_2. Therefore, it is wasteful to use both w_1 and w_2 in the same frequency in E. One efficient way to benefit from the existence of two words that compete on E is by partitioning E into subsets X and Y, and use u_1 more frequently in X, and u_2 more frequently in Y. Thus, when u_1 or u_2 is uttered, it is easier to retrieve the intended sub-environment, since it is more likely to be X or Y, respectively.

For example, the principle of contrast predicts such interaction between *hence* and *therefore*. Originally, *hence* was a locative used to indicate place of origin ("from hence"). An increase in the DM use of *hence* would lead to competition with *therefore* on the sentence-initial DM environment ("hence, John is smart"), so this environment would be partitioned into sentence-initial DM and mid-sentence DM ("John is therefore smart"), such that *hence* is preferred in the former, and *therefore* is preferred in the latter. This way, speakers can save cognitive effort when choosing among DMs which compete on the same meaning - namely, justification. This process is illustrated in Table 1. Refer to Section 4 for the actual distribution of *hence* and *therefore* in this environment.

The relaxed version of the principle of contrast predicts that when a new word immigrates to an environment E, speakers will be motivated to use it with different probabilities than other words that can appear in E without a change in meaning. This entails that when a word is introduced into a new environment, it will lead to words that mean essentially the same in that environment to become increasingly less frequent in that environment over time. This is the prediction tested in this paper.

3 Methodology

3.1 Setup

To test the proposal made in Section 2.2, I computed the co-distributions of groups of words that compete on related uses in the Penn Parsed Corpora of Early Modern English (Kroch and Delfs, 2004) and the Parsed Corpora of Early English Correspondence (Taylor and Nevalainen, 2006). Words chosen for this study were ones that had two or more distinct annotation schemes in the corpus. Each annotation scheme is treated as a separate use of the word. All distinguishable uses of each highlighted word were identified throughout the corpus. For example, the contrast use of *but* was annotated as a conjunction, while the exception use was annotated as a preposition. For each comparison set of words W which compete on some environment E, the pattern which defines E was chosen to be the weakest possible regular expression over tree structures which captures exactly one of the uses of each $w \in W$.

As was discussed in Section 1, the advantage of this approach is that (i) it enables us to distinguish different uses of one word, and (ii) it yields robust predictions about function words, taking advantage of the information contained in tree structures.

I looked for environments that can be characterized syntactically, in which more than one word can appear without a substantial change in meaning. This enables us to automatically capture groups of words that compete on the same environment. However, the ability to distinguish different uses of function words comes at the cost of limited scope. The only words that can be examined are ones that satisfy the above restrictions. This might introduce some statistical bias, since the examined words are not based on a random sample. However, this bias was traded off in exchange for higher precision and ability to distinguish multiple senses of function words, as explained above. As a point of comparison, bag-of-words based word vectors cannot achieve this level of precision, since (i) they are inherently monosemous and (ii) word vectors for function words are highly uninformative relatively to content words, since function words are frequent nearly everywhere (Section 4.1).[2]

The words selected for this study are *very*, *thus*, *but*, *except*, *though*, *therefore*, *still*, *yet*, *from*, *hence*, *as* and *when*.

[2] There exist vector space models that were trained on syntactically annotated corpora (MacAvaney and Zeldes, 2018; Levy and Goldberg, 2014; Komninos and Manandhar, 2016), which might address point (ii) above, but it would be difficult to apply such models for a historical study since dependency parsers were largely trained on Modern English data, and therefore cannot be used to annotate historical texts. Hence, in this study, manually annotated corpora were used.

	Loc.	Init.	Mid.
Hence	.9	.1	0
Therefore	0	.9	.1

(a) Before increase

Loc.	Init.	Mid.
.7	.3	0
0	.9	.1

(b) Increase

Loc.	Init.	Mid.
.3	.7	0
0	.2	.8

(c) After increase

Table 1: Illustration of hypothetical proposed interaction between words. A hypothetical distribution of *therefore* and *hence* over environments (locative, sentence-initial justification DM, mid-sentence justification DM) is taken as an example. The rise in sentence initial uses of *hence* creates competition, and this competition is resolved by *therefore* becoming less frequent in that environment.

3.2 Competing Pairs

Still and *yet* compete on their positive polarity use (denoted by the variables *_adv_pos, demonstrated in (1)). Additionally, they compete on an adverbial use (*_adv) following a raised clause introduced by a complementizer/preposition, as demonstrated in (2).

(1) And consequently, they may still with greater ease begin with it, ...

(2) If I can come again, we are still to have our ball.

Very and *thus* compete on an intensifier degree-adverbial use (*_adv_deg). For *very*, this is the only use, but *thus* has 3 syntactically distinguishable uses, which I term as follows: degree-adverbial (3) (modifying an adjective), manner adverbial (4) (modifying a verb), and discourse particle (5) (modifying a clause). Generally speaking, the degree-use can be paraphrased as ‘to that extent’; the manner-use can be paraphrased as ‘in that way’; the discourse particle use can be paraphrased as ‘for that reason’.

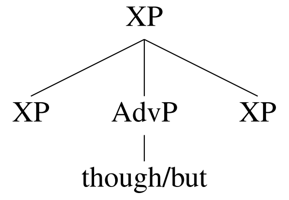
(3) We are, however, thus little acquainted with...

(4) I wished when I heard them say thus, that...

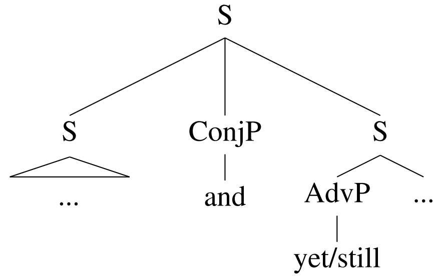
(5) And thus I bid you farewell from my house at foston this ix of november.

But competes with *except* on the exception use (e.g. ”all but a few”). This use is marked as a preposition, and therefore it is denoted by the variable *_p. *Though* and *but* are two contrast words that compete on the contrast coordination environment *_conj. The structure corresponding to this use is demonstrated in (6).

(6)



(7)



Hence competes with *therefore* on the sentence-initial discourse particle environment (*_dp_top). The locative use of *hence* hence_loc (meaning ‘from here’), competes with *from*. *From* is measured by absolute frequency since its only use is the locative one.

As and *when* can both introduce temporal clauses, e.g. ‘as/when you arrive’. In the Penn Corpora, temporal complementizers are tagged as prepositions, and therefore I use the pattern *_p to capture these uses.

3.3 Statistical Model

To investigate the hypothesis stated in Section 2.2, relative and absolute frequency counts were collected for each use of each word, and the counts for each competing pair were compared to each other over time. Formally, Each use pattern p of word w was represented as a vector of frequencies over the set T of all 50-year intervals between 1150 CA and 1950 CA (closed to the left and open to the right). For some $\tau \in T$, $\overrightarrow{\tau}$ is the vector of intervals that are greater or equal to τ. For example, $\overrightarrow{[1800, 1850)}$ is the vector $([1800, 1850), [1850, 1900), [1900, 1950))$. Each use u is represented as a random variable U, such that $U_{\overrightarrow{\tau}}$ is the vector of the count, for each interval $\tau' \geq \tau$ (Read: ”following or equal to τ”) of matches for the formal pattern that corresponds to

u. For example, the random variable $\mathtt{but}_{cont}, \overrightarrow{\tau}$ ($cont$ for contrast) is the vector of counts of all matches of the formal pattern [$_{\text{conj}}$ *but*], that is, all sub-parses that contain a *but* and are directly dominated by a *conj* node, starting from interval τ and onward. $W_{i,\overrightarrow{\tau}}$ is the variable corresponding to the vector of counts of the i-th word in τ and the following intervals.

The absolute frequency of a variable $U_{\overrightarrow{\tau}}$ is defined as the vector that, for each $s \geq t$, stores in its k-th position the value of $U_{\overrightarrow{i}}$ at position k divided by the counts of all V words in the vocabulary at position k:

$$\frac{U_{\overrightarrow{\tau}}}{\sum_{i=1}^{|V|} W_{i,\overrightarrow{\tau}}}$$

If a word has only one relevant use, then its absolute frequency was used as its frequency measure. For example, the word *very* had only one use that was investigated, namely, its adverbial use. The relative frequency of some use of the i-th word, with variable $U_{\overrightarrow{\tau}}$, is defined as the proportion of the U values with respect to the W_i values for each interval:

$$\frac{U_{\overrightarrow{\tau}}}{W_{i,\overrightarrow{\tau}}}$$

If a word had more than one use, its relative frequency was used as its frequency measure.I use $F(U_{\overrightarrow{\tau}})$ to denote the frequency measure of $U_{\overrightarrow{\tau}}$.

The hypothesis states that for each pair of uses competing on an environment, there exists a point in time t such that one of the uses becomes more frequent following t and the other becomes less frequent following t.[3] That is, for each comparison pair CP, there are $U, V \in CS$ such that there exists some $\tau \in T$ such that $Cov(F(U_{\overrightarrow{\tau}}), T_{\overrightarrow{\tau}}) <$ o and $Cov(F(V_{\overrightarrow{\tau}}), T_{\overrightarrow{\tau}}) >$ o. Conceptually, V corresponds to the newer uses which push the older uses, U, out of the environment of competition.

This entails that the trends (i.e. true population regression lines) for U and V in the interval $T_{\overrightarrow{\tau}}$ have opposite slopes, which means that they cross at some interval τ' (which may or may not be in T). This means that there is a point in time (τ') starting from which, one use's frequency grows over time, while the other use's frequency decreases over time. Formally, we have that:

$$|F(U_{\tau'}) - F(V_{\tau'})| = O((\tau - \tau')^2)$$

following the definition of O complexity. $\tau - \tau'$ is the difference between the beginning of τ' and the end of τ. In other words, the hypothesis predicts that the difference between the frequencies of the words in each pair grow quadratically as a function of the distance from the point in time in which the trends cross each other.

To model this behavior, a cubic model was fitted to the differences between each pair as a function of T. To account for the fact that the interaction might only take place in a subinterval of T, the model had two splines, one at 1300 and one at 1650.[4] This allows for a coefficient change at those points, which reflects the fact that the effect between the two words might be different for different subintervals of T. The hypothesis predicts that such a model would have a significant fit at $\alpha = .05$.

To verify that the use frequencies indeed change as a linear function of time, for each use U it was tested whether there exists an interval $\tau \in T$ such that the linear model:

$$F(U_{\overrightarrow{\tau}}) \sim \overrightarrow{\tau}$$

has a significant slope coefficient at $\alpha = .05$. The existence of such a trend shows that U is not constant in time, which entails that if the cubic model is significant, then the two trends have opposite signs (i.e. they are crossing).

4 Results

Coefficients and significance levels for all comparison sets are displayed in Table 2. Each model has 3 coefficients, since the 2 knots partition the intervals into 3 parts. Frequency differences by century are plotted along with model curves. All models were significant, with the exception of the model for *thus* and *very*. All words, with the exception of *still*, were found to change as a linear function of time starting from some year, as described above.

The local extrema of the curves indicate the points at which, according to the hypothesis, the

[3] Following the discussion in Section 2.2, the use that becomes less frequent is likely to be the older one, but the hypothesis does not require for this to be the case. The reason is that semantic change can also make words less abstract. For example, *computer* used to denote any computing device, but now typically denotes any computing device which is not tablet-shaped.

[4] As is common practice, the splines were placed at the quantiles of the x axis (rounded).

Env.	Use 1	Use 2	Coef1	Coef2	Coef3	Signif. level
Temporal Adv.	`still_adv`	`yet_adv_pos`	0.83	1.35	0.45	**
Degree/Manner	`very`	`thus_adv_deg`	-0.16	-0.06	-0.0072	.
Justification	`hence_dp`	`therefore_dp_top`	-0.1	-0.73	1.22	***
Locative	`hence_adv`	`from`	1.04	1.56	-0.43	***
Exception	`but_p`	`except`	2.06	0.65	-1e-04	***
Contrast	`though`	`but_contrast`	-3.53	-1.62	-1.42	**
Temporal Comp.	`as_p`	`when_p`	0.1	-1.03	0.12	*

Table 2: Trends by use. Slopes are measured starting from the century the trend started. Signif. codes: 0.001 ** 0.01 * 0.05 '.' .1 ' ' 1

trend lines of the two words cross each other. Due to the X shape formed at those junction points, the absolute difference between the two trends grows cubically around them, which leads to the cubic fit. Curvature change at the splines indicate that the trends have shifted at those points. The chronologically latest trend is indicated by the rightmost parabola, which is the one we are interested in.

For each plot, the order in which the word pair was written reflects the order the subtraction operation applied to the two words' frequencies. Thus, a convex (concave) parabola indicates that the first (second) word's regression line (i.e. its sample trend) has a positive slope while the first (second) word's regression line has a negative slope.

These results suggest that *but*'s contrast use increases at the expense of *though*, and at the same time *except* pushes *but* out of the exception use. *From* pushes *hence* out of the locative use, and *hence* pushes *therefore* out of the justification use. *As* pushes *when* out of the temporal complementizer use. *Yet* seems to interact with *still* in the same way in the positive polarity use, but a definitive linear trend for *still* was failed to be established, so it may be that the difference observed between *stil* and *yet* is only due to a change in *yet* and not a change in *still*. The results also somewhat support the idea that *very* pushes *thus* out of the degree/manner environment, but the model's level of significance warrants further investigation.

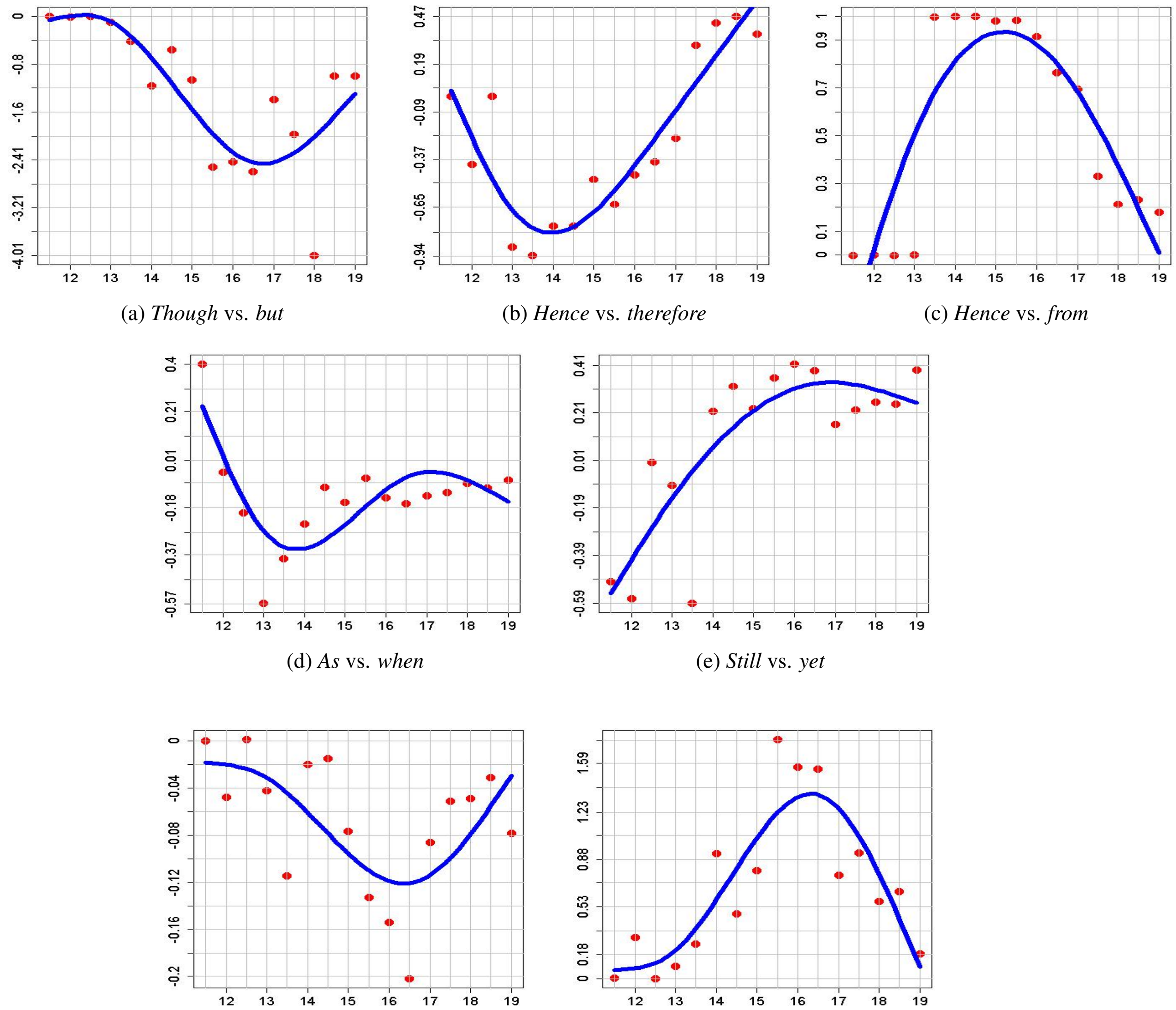
4.1 Word Cooccurrence

The proposal made in this paper concerns changes in the structural distribution of semantically similar function words. Function words that share one or more uses are claimed to diverge over time in their syntactic distribution. That is, the syntactic positions they occur in will distribute differently from each other. This claim does not, however, predict that they will occur near different *words*. As Xu and Kemp (2015) report, there is no evidence that semantic similarity leads to difference in word cooccurrence.

Xu and Kemp's hypothesis was formulated in terms of word2vec models (Mikolov et al., 2013), which approximate high-order functions of word cooccurrence. The principle of contrast was translated into the hypothesis that over time, similar words will diverge more than control words in terms of the cosine distance of their vectors. This hypothesis, which was falsified, postulates that two similar words will over time cooccur with different words.

Thus, Xu and Kemp's results suggest that semantically similar words do not tend to diverge in their cooccurrence patterns. They do not, however, exclude the possibility that semantically similar words tend to diverge in the structural positions they assume. This latter possibility is the thesis advocated in this paper.

To verify that the information contained in parsed structures is not fully recoverable from word cooccurrence, I divided the dataset roughly into the Early Modern and Late Modern periods, and for each period, I collected raw cooccurrence matrices for each word that occurred at least twice, with window of size 20. This yields for each word w a vector representation in which the ith position stores the number of times w occurred with the ith vocabulary item in the same sentence. Due to the relatively small size of the dataset, it is not suitable for learning higher order vector representations of words as in word2vec, since such models require larger amounts of data in order to generalize properly. For each pair of words in the vocabulary I computed the change between the early and late periods:



(a) *Though* vs. *but*

(b) *Hence* vs. *therefore*

(c) *Hence* vs. *from*

(d) *As* vs. *when*

(e) *Still* vs. *yet*

(f) *Very* vs. *thus*

(g) *But* vs. *except*

Figure 1: Fitted curves by use pair. The horizontal axis shows centuries and the vertical axis shows the difference between the frequencies of the uses.

$$\delta(w_1, w_2) = |\cos_E(w_1, w_2) - \cos_L(w_1, w_2)|$$

where $\cos_{E/L}$ is the cosine angle between the co-occurrence vectors for w_1 and w_2 for the early and late matrices, respectively. This quantity represents the change in similarities between w_1 w_2 from E to L.

I then considered $\mathbb{E}[\delta(w_1, w_2)]$ for each pair w_1, w_2 of highlighted words as partitioned above (*however-so*, *still-yet*, etc'), and compared it to the $\mathbb{E}[\delta w_1, w_2]$ for every other pair w_1, w_2 of words. This experiment was performed with two cutoff points between the early and late periods: 1700 and 1755 (1755 is the year in which the Dictionary of the English Language, was published which standardized spelling and vocabulary, but 1700 gives a more balanced partition in terms of quantity). For both cutoff points, the two means were different at $\alpha = .01$. The mean for the highlighted pairs was around .1, and the mean for the other pairs was around .33.

This result suggests that no distributional difference is observed between semantically similar word pairs based on word cooccurrence alone. This is in line with (Xu and Kemp, 2015), according to which semantically similar word pairs do not tend to diverge in their cooccurrence patterns over time.

5 Discussion and Conclusions

This paper examined the hypothesis that when a new word immigrates to a new environment, older alternatives tend to decrease in frequency in that environment. Results show that the hypothesis was validated in all cases except `still_adv_pp`. Notice that, in many cases, the r^2 statistic was relatively small. This is unsurprising, since the hypothesis only accounts for a small amount of variation in the data. In other words, time is not the only variable that affects the frequency of different uses of a word.

These results support the proposal detailed in Section 2, namely that as words become more abstract, they compete with old words that share their new environment, leading to the old words being driven to new environments. The results are in alignment with the literature on contrast (Section 2.1). Additionally, it has been shown that these results are not replicated when considering word cooccurrence alone, which suggests that the effects observed are indeed due to structural differences between different uses of the same word. These is fundamentally different from the way content words change, because as has been shown in previous studies (Section 1) content words often move to new distributional environments (in terms of cooccurrence) without any change in syntactic position.

The methodology applied in this study - of using hand-crafted syntactic patterns to distinguish between different uses of the same word - allows for a detailed examination of specific word pairs, which enables us to test highly refined hypotheses. However, this precision is traded off for empirical limitedness, as only a closed set of words satisfies the conditions necessary to be distinguishable in an annotated corpus.

A major limitation of the current study is its narrow empirical coverage, as the study examines a closed set of word pairs. A desirable extension would be testing the same hypothesis across the board for the entire vocabulary. Such an extension would require an innovative method for automatically detecting environments on which word pairs compete.

This phenomenon may be viewed as part of speakers' efforts to maximize utility by conveying the greatest amount of information with the least amount of cognitive effort (as discussed in Section 2). Under this assumption, speakers rely on each other's rational faculties to infer the most likely interpretation of an utterance. lexical meaning is subject to pragmatic considerations of conveying the greatest amount of detail with the least amount of effort. This may explain some of the findings in this study, considering the economic benefit of dividing the labor between different words. These motivations were illustrated in Section 2.

The novelty of the data presented here compared to previous approaches is (i) the application to functional words, specifically prepositions, DMs, and functional adverbs (e.g. *very*, *so*, *really*), and (ii) the ability to compare different uses of the same word. For example, consider the case of *still* vs. *yet*. Figure 1e shows an increase in the positive-polarity and adverbial uses of *still*, followed by a decrease in those same uses of *yet*. This suggests that, following *still*'s immigrating to environments which happen to be shared with *yet*, due to the principle of contrast, those same uses of *yet* are dispreferred, leading to their decrease. This results in the predictions spelled out in Section 2.

Interestingly, note that none of the uses explored in this paper disappear completely. This result is surprising if one considers a naive interpretation of RSA or other game-theoretic approaches to semantic change such as Ahern and Clark (2017). Based on such approaches, one might expect that novel competitors on an environment would eliminate older ones completely, since storing their use in that environment requires unjustified cognitive effort. A possible direction for further research is extending such models to account for the existence of two words that share a use by exploring which distinctions they do mark within that use. A possible explanation is that they are used to mark meta-linguistic information about the utterance, such as the sociology or attitude of the speaker, but this question is left for future experiments.

Acknowledgements

I would like to thank Chris Potts, Paul Kiparsky and Dan Jurafsky for invaluable advise and comments on several drafts of this paper. Their input was essential to the completion of this project. I would also like to thank the Howard B. Kerzner Graduate Fellowship which funds my PhD studies at Stanford University and allowed me access to the resources necessary to complete this project. Additionally, I thank the anonymous reviewers of

this workshop who provided detailed comments on the paper.

References

Christopher Ahern and Robin Clark. 2017. Conflict, cheap talk, and Jespersen's cycle. *Semantics and Pragmatics*, 10.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. 36(4):673–721.

Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. In *Proceedings of Third Italian Conference on Computational Linguistics*.

Michel Bréal. 1897. *Essai de sémantique:(science des significations)*. Paris: Hachette.

Eve V Clark. 1990. On the pragmatics of contrast. *Journal of child language*, 17(02):417–431.

Eve V Clark and Herbert H Clark. 1979. When nouns surface as verbs. *Language*, pages 767–811.

Kris De Jaegher and Robert van Rooij. 2014. Game-theoretic pragmatics under conflicting and common interests. *Erkenntnis*, 79(4):769–820.

Ferdinand De Saussure. 1916. *Cours de linguistique générale: Édition critique*. Lausanne; Paris: Payot.

Ashwini Deo. 2015. The semantic and pragmatic underpinnings of grammaticalization paths: The progressive to imperfective shift. *Semantics and Pragmatics*, 8(14):1–52.

Dankmar Enke, Roland Muhlenbernd, and Igor Yanovich. 2016. The emergence of the progressive to imperfective diachronic cycle in reinforcement-learning agents. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11). New Orleans*.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(3):998–998.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modelling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Herbert P. Grice. 1975. Logic and conversation. In *Syntax and semantics (Vol. 3)*, pages 41–58.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. *Proceedings of EMNLP*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of ACL*.

Jerry R Hobbs. 1985. *On the coherence and structure of discourse*. Research report, Center for the Study of Language and Information, Stanford University.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1490–1500.

Beatrice Santorini Kroch, Anthony and Lauren Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). CD-ROM, first edition, release 3. Department of Linguistics, University of Pennsylvania.

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. ACM.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.

Sean MacAvaney and Amir Zeldes. 2018. A deeper look into dependency-based word embeddings. *arXiv preprint arXiv:1804.05972*.

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

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. 33(2):161–199.

Hermann Paul. 1898. *Prinzipien der sprachgeschichte*. Halle: Max Niemeyer Verlag.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.

Gerhard Schaden. 2012. Modelling the aoristic drift of the present perfect as inflation an essay in historical pragmatics. *International Review of Pragmatics*, 4(2):261–292.

Eve Sweetser. 1991. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Arja Nurmi Anthony Warner Susan Pintzuk Taylor, Ann and Terttu Nevalainen. 2006. The York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC). Oxford Text Archive, first edition. Department of Linguistics, University of York.

Elizabeth C. Traugott and Richard B. Dasher. 2002. *Regularities in semantic change*. Cambridge: Cambridge University Press.

Elizabeth Closs Traugott. 1995. The role of the development of discourse markers in a theory of grammaticalization. *ICHL XII, Manchester*, 123.

Elizabeth Closs Traugott and John Waterhouse. 1969. 'already' and 'yet': a suppletive set of aspect-markers? *Journal of Linguistics*, 5(02):287–304.

Stephan Ullman. 1962. *Semantics: An Introduction to the Science of Meaning*. New York City: Barnes & Noble.

Thomas Wasow. 2015. Ambiguity avoidance is overrated. *Ambiguity: Language and Communication*, page 29.

David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: a theory of composition for distributional semantics. *Computational Linguistics, special issue on Formal Distributional Semantics*, 42(4):727–761.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proc. Annual Conf. of the Cognitive Science Society*.

Igor Yanovich. Analyzing imperfective games. *lingbuzz/002652*.