Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony

Florent Perek
Princeton University
Princeton, NJ, USA
fperek@princeton.edu

Abstract

This paper describes an application of distributional semantics to the study of syntactic productivity in diachrony, i.e., the property of grammatical constructions to attract new lexical items over time. By providing an empirical measure of semantic similarity between words derived from lexical co-occurrences, distributional semantics not only reliably captures how the verbs in the distribution of a construction are related, but also enables the use of visualization techniques and statistical modeling to analyze the semantic development of a construction over time and identify the semantic determinants of syntactic productivity in naturally occurring data.

1 Introduction

Language change does not exclusively consist of drastic shifts in 'core' aspects of grammar, such as changes in word order. Variation in usage, which can occur in no more than a few decades, is much more common, and to many linguists constitutes linguistic change in the making. Among these aspects of language use that are subject to diachronic change, this paper is concerned with the productivity of syntactic constructions, i.e., the range of lexical items with which a construction can be used. A given construction might occur with very different distributions at different points in time, even when the function it conveys remains the same. This is what Israel (1996) finds for the pattern "Verb *one's way* Path", commonly called the *way*-construction (Goldberg, 1995), exemplified by (1) and (2) below.

(1) They hacked their way through the jungle.
(2) She typed her way to a promotion.

As reported by Israel, examples like (1), in which the main verb describes the physical means whereby motion towards a goal is enabled, are attested as early as the 16th century, but it was not until the 19th century that examples like (2) started to appear, in which the action depicted by the verb provides a more indirect (and abstract) way of attaining the agent's goal.

The productivity of a construction may appear partly arbitrary, but a growing body of evidence suggests that it is tied to the previous experience of speakers with that construction (Barðdal, 2008; Bybee and Eddington, 2006; Suttle and Goldberg, 2011). More specifically, previous research points to a strong semantic component, in that the possibility of a novel use depends on how it semantically relates to prior usage. Along these lines, Suttle and Goldberg (2011, 1254) posit a criterion of coverage, defined as "the degree to which attested instances 'cover' the category determined jointly by attested instances together with the target coinage". Coverage relates to how the semantic domain of a construction is populated in the vicinity of a given target coinage, and in particular to the density of the semantic space.

The importance of semantics for syntactic productivity implies that the meaning of lexical items must be appropriately taken into account when studying the distribution of constructions, which calls for an empirical operationalization of semantics. Most existing studies rely either on the semantic intuitions of the analyst, or on semantic norming studies (Bybee and Eddington, 2006). In this paper, I present a third alternative that takes advantage of advances in computational linguistics and draws on a distributionally-based measure of semantic similarity. On the basis of a case study of the construction "V *the hell out of* NP", I show how distributional semantics can profitably be applied to the study of syntactic productivity.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 309–314,
Baltimore, Maryland, USA, June 23-25 2014.

2 The *hell*-construction

The case study presented in this paper considers the syntactic pattern "V *the hell out of* NP", as exemplified by the following sentences from the Corpus of Contemporary American English (COCA; Davies, 2008):

(3) Snakes just scare the hell out of me.
(4) It surprised the hell out of me when I heard what he's been accused of.
(5) You might kick the hell out of me like you did that doctor.

The construction generally conveys an intensifying function (very broadly defined). Thus, *scare/surprise the hell out of* means "scare/surprise very much", and *kick the hell out of* means "kick very hard". The particular aspect that is intensified may be highly specific to the verb and depend to some extent on the context. *Scare* and *beat* are the most typical verbs in that construction (and arguably the two that first come to mind), but a wide and diverse range of other verbs can also be found, such that *avoid* in (6), *drive* (a car) in (7) and even an intransitive verb (*listen*) in (8):

(6) I [...] avoided the hell out of his presence.
(7) But you drove the hell out of it!
(8) I've been listening the hell out of your tape.

To examine how the construction evolved over time, I used diachronic data from the Corpus of Historical American English (COHA; Davies 2010), which contains about 20 million words of written American English for each decade between 1810 and 2009 roughly balanced for genre (fiction, magazines, newspapers, non-fiction). Instances of the *hell*-construction were filtered out manually from the results of the query "[v*] the hell out of", mostly ruling out locative constructions like *get the hell out of here*. The diachronic evolution of the verb slot in terms of token and type frequency is plotted in Figure 1. Since the corpus size varies slightly in each decade, the token frequencies are normalized per million words.

The construction is first attested in the corpus in the 1930s. Since then, it has been steadily increasing in token frequency (to the exception of a sudden decrease in the 1990s). Also, more and more different verbs are attested in the construction, as shown by the increase in type frequency.

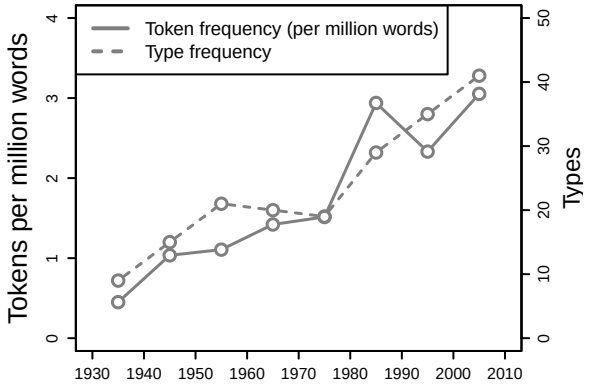

Figure 1: Diachronic development of the *hell*-construction in terms of normalized token frequency and type frequency

This reflects a general expansion of the productivity of the construction, but it does not show what this productivity consists of. For instance, it does not say what kinds of verbs joined the distribution and to what extent the distribution becomes semantically more diverse over time. To answer these questions, I will analyze the distribution of the construction from a semantic point of view by using a measure of semantic similarity derived from distributional information.

3 Distributional measure of semantic similarity

Drawing on the observation that words occurring in similar contexts tend to have related meanings (Miller and Charles, 1991), distributional approaches to semantics seek to capture the meaning of words through their distribution in large text corpora (Lenci, 2008; Turney and Pantel, 2010; Erk, 2012). One benefit of the distributional semantics approach is that it allows semantic similarity between words to be quantified by measuring the similarity in their distribution. This is achieved by means of a vector-space model that assigns an array of numerical values (i.e., a vector) derived from distributional information to each word. A wide range of distributional information can be employed in vector-based models; the present study uses the 'bag of words' approach, which is based on the frequency of co-occurrence of words within a given context window. According to Sahlgren (2008), this kind of model captures to what extent words can be substituted for each other, which is a good measure of semantic similarity between verbs. As it turns out, even this

relatively coarse model captures semantic distinctions in the distribution of the *hell*-construction that make intuitive sense.

All instances of the relevant verbs were extracted from the COCA[1] with their context of occurrence. In order to make sure that enough distributional information is available to reliably assess semantic similarity, verbs with less than 2,000 occurrences were excluded, which left 92 usable items (out of 105). The words in the sentence contexts extracted from the COCA were lemmatized and annotated for part-of-speech using TreeTagger (Schmid, 1994). The part-of-speech annotated lemma of each collocate within a 5-word window was extracted from the COCA data to build the co-occurrence matrix recording the frequency of co-occurrence of each verb with its collocates. Only the nouns, verbs, adjectives, and adverbs listed among the 5,000 most frequent words in the corpus were considered (to the exclusion of *be*, *have*, and *do*), thus ignoring function words (articles, prepositions, conjunctions, etc.) and all words that did not make the top 5,000.

The co-occurrence matrix was transformed by applying a Point-wise Mutual Information weighting scheme, using the DISSECT toolkit (Dinu et al., 2013), to turn the raw frequencies into weights that reflect how distinctive a collocate is for a given target word with respect to the other target words under consideration. The resulting matrix, which contains the distributional information (in 4,683 columns) for 92 verbs occurring in the *hell*-construction, constitutes the semantic space under consideration in this case study. Pairwise distances between the target verbs were calculated using the cosine distance. The rest of the analysis was conducted on the basis of this distance matrix in the R environment (R Development Core Team, 2013).

[1]The COCA contains 464 million words of American English consisting of the same amount of spoken, fiction, magazine, newspaper, and academic prose data for each year between 1990 and 2012. Admittedly, a more ecologically valid choice would have been to use data from a particular time frame to build a vector-space model for the same time frame, but even the twenty-odd million words per decade of the COHA did not prove sufficient to achieve that purpose. This is, however, not as problematic as it might sound, since the meaning of the verbs under consideration are not likely to have changed considerably within the time frame of this study. Besides, using the same data presents the advantage that the distribution is modeled with the same semantic space in all time periods, which makes it easier to visualize changes.

4 Application of the vector-space model

4.1 Semantic plots

One of the advantages conferred by the quantification of semantic similarity is that lexical items can be precisely considered in relation to each other, and by aggregating the similarity information for all items in the distribution, we can produce a visual representation of the structure of the semantic domain of the construction in order to observe how verbs in that domain are related to each other, and to immediately identify the regions of the semantic space that are densely populated (with tight clusters of verbs), and those that are more sparsely populated (fewer and/or more scattered verbs). Multidimensional scaling (MDS) provides a way both to aggregate similarity information and to represent it visually. This technique aims to place objects in a space with two (or more) dimensions such that the between-object distances are preserved as much as possible.

The pairwise distances between verbs were submitted to multidimensional scaling into two dimensions.[2] To visualize the semantic development of the *hell*-construction over time, the diachronic data was divided into four successive twenty-year periods: 1930-1949, 1950-1969, 1970-1989, and 1990-2009. The semantic plots corresponding to the distribution of the construction in each period are presented in Figure 2. For convenience and ease of visualization, the verbs are color-coded according to four broad semantic groupings that were identified inductively by means of hierarchical clustering (using Ward's criterion).[3]

By comparing the plots in Figure 2, we can follow the semantic development of the *hell*-construction. The construction is strikingly centered around two kinds of verbs: mental verbs (in red: *surprise*, *please*, *scare*, etc.) and verbs of hitting (most verbs in green: *smash*, *kick*, *whack*, etc.), a group that is orbited by other kinds of forceful actions (such as *pinch*, *push*, and *tear*). These two types of verbs account for most of the distribution at the onset, and they continue to

[2]Non-metric MDS was employed (Kruskal, 1964), using the function `isoMDS` from the R package MASS.

[3]Another benefit of combining clustering and MDS stems from the fact that the latter often distorts the data when fitting the objects into two dimensions, in that some objects might have to be slightly misplaced if not all distance relations can be simultaneously complied with. Since cluster analysis operates with all 4,683 dimensions of the distributional space, it is more reliable than MDS, although it lacks the visual appeal of the latter.

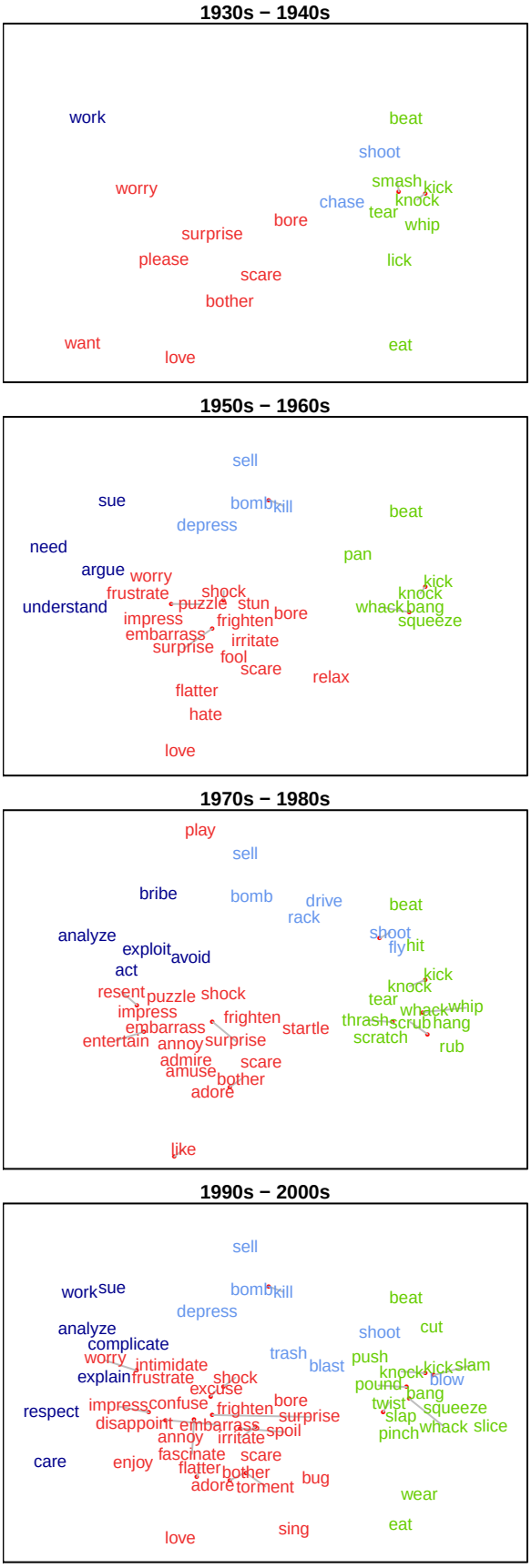

Figure 2: Semantic plots of the *hell*-construction in four time periods.

weigh heavily throughout the history of the construction. These two classes also correspond to the regions of the semantic domain that attract the most new members, and they constantly do so in all periods. Outside of these two clusters, the semantic space is much more sparsely populated. In the first period (1930-1949), only a few peripheral members are found. They are joined by other distantly related items in later periods, although by no more than a handful in each. In other words, the construction is markedly less productive in these outer domains, which never form proper clusters of verbs.

In sum, the semantic plots show that densely populated regions of the semantic space appear to be the most likely to attract new members. Outside of the two identified domains of predilection, other classes never become important, assumedly because they do not receive a "critical mass" of items, and therefore attract new members more slowly.

4.2 Statistical analysis

With the quantification of semantic similarity provided by the distributional semantic model, it is also possible to properly test the hypothesis that productivity is tied to the structure of the semantic space. On the reasonable assumption that the semantic contribution of the construction did not change, and therefore that all verbs ever attested in it are equally plausible from a semantic point of view, the fact that some verbs joined the distribution later than others is in want of an explanation. In view of the observations collected on the semantic plots and in line with previous research (especially Suttle and Goldberg's notion of coverage), I suggest that the occurrence of a new item in the construction in a given period is related to the density of the semantic space around that item in the previous period. If the semantic space around the novel item is dense, i.e., if there is a high number of similar items, the coinage will be very likely. The sparser the semantic space around a given item, the less likely this item can be used.

The measure of density used in this study considers the set of the N nearest neighbors of a given item in the semantic space, and is defined by the following formula:

$$Density_{V,N} = 1 - \frac{\sum_{n=1}^{N} d(V, V_n)}{N}$$

where $d(V, V_n)$ is the distance between a verb V

and its n^{th} nearest neighbor. In plain language, density equals one minus the mean distance to the N nearest neighbors. The latter value decreases with space density (i.e., if there are many close neighbors), and is therefore technically a measure of sparsity; since cosine distances are between 0 and 1, subtracting the mean distance from one returns a measure of density within the same boundaries.

This measure of density was used as a factor in logistic regression to predict the first occurrence of a verb in the construction, coded as the binary variable OCCURRENCE, set to 1 for the first period in which the verb is attested in the construction, and to 0 for all preceding periods (later periods were discarded). For each VERB-PERIOD-OCCURRENCE triplet, the density of the semantic space around the verb in the immediately preceding period was calculated. Six different versions of the density measure, with the number of neighbors under consideration (N) varying between 3 and 8, were used to fit six mixed effects regression models with OCCURRENCE as the dependent variable, DENSITY as a fixed effect, and random by-verb intercepts and slopes (Bates et al., 2011). The results of these models are summarized in Table 1.

N	Effect of DENSITY	*p*-value
3	0.7211	0.195
4	0.8836	0.135
5	1.0487	0.091 (.)
6	1.2367	0.056 (.)
7	1.4219	0.034 (*)
8	1.6625	0.017 (*)

Table 1: Summary of logistic regression results for different values of N. Model formula: OCCURRENCE $\sim$ DENSITY + (1 + DENSITY|VERB). Marginally significant effects are marked with a period (.), significant effects with a star (*).

For all values of N, we find a positive effect of DENSITY, i.e., there is a positive relation between the measure of density and the probability of first occurrence of a verb in the construction. However, the effect is only significant for $N \geq 7$; hence, the hypothesis that space density increases the odds of a coinage occurs in the construction is supported for measures of density based on these values of N.

More generally, the *p*-value decreases as N increases, which means that the positive relation between DENSITY and OCCURRENCE is less systematic when DENSITY is measured with fewer neighbors. This is arguably because a higher N helps to better discriminate between dense clusters where all items are close together from looser ones that consist of a few 'core' items surrounded by more distant neighbors. This result illustrates the role of type frequency in syntactic productivity: a measure of density that is supported by a higher number of types makes better prediction than a measure supported by fewer types. This means that productivity not only hinges on how the existing semantic space relates to the novel item, it also occurs more reliably when this relation is attested by more items. These finding support the view that semantic density and type frequency, while they both positively influence syntactic productivity, do so in different ways: density defines the necessary conditions for a new coinage to occur, while type frequency increases the confidence that this coinage is indeed possible.

5 Conclusion

This paper reports the first attempt at using a distributional measure of semantic similarity derived from a vector-space model for the study of syntactic productivity in diachrony. On the basis of a case study of the construction "V *the hell out of* NP" from 1930 to 2009, the advantages of this approach were demonstrated. Not only does distributional semantics provide an empirically-based measure of semantic similarity that appropriately captures semantic distinctions, it also enables the use of methods for which quantification is necessary, such as data visualization and statistical analysis. Using multidimensional scaling and logistic regression, it was shown that the occurrence of new items throughout the history of the construction can be predicted by the density of the semantic space in the neighborhood of these items in prior usage. In conclusion, this work opens new perspectives for the study of syntactic productivity in line with the growing synergy between computational linguistics and other fields.

References

Johana Barðdal. 2008. *Productivity: Evidence from Case and Argument Structure in Icelandic*. John Benjamins, Amsterdam.

Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker. 2011. *lme4: Linear mixed-effects models using S4 classes. R package*. URL: http://CRAN.R-project.org/package=lme4

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Joan Bybee and David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2):323–355.

Mark Davies. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/

Mark Davies. 2010. *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/

Georgiana Dinu, The Nghia Pham and Marco Baroni. 2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Michael Israel. 1996. The way constructions grow. In Adele E. Goldberg (ed.), *Conceptual structure, discourse and language*, pages 217–230. CSLI Publications, Stanford, CA.

Joseph Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20(1):1–31.

George Miller and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

R Development Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna; URL: http://www.R-project.org/

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

Laura Suttle and Adele Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics*, 49(6):1237–1269.

Peter Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.