Using distributional semantics to study syntactic productivity in diachrony: A case study

Perek, Florent

DOI:
10.1515/ling-2015-0043

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Perek, F 2016, 'Using distributional semantics to study syntactic productivity in diachrony: A case study', *Linguistics*, vol. 54, no. 1, pp. 149-188. https://doi.org/10.1515/ling-2015-0043

Florent Perek*

Using distributional semantics to study syntactic productivity in diachrony: A case study

DOI 10.1515/ling-2015-0043

Abstract: This paper investigates syntactic productivity in diachrony with a data-driven approach. Previous research indicates that syntactic productivity (the property of grammatical constructions to attract new lexical fillers) is largely driven by semantics, which calls for an operationalization of lexical meaning in the context of empirical studies. It is suggested that distributional semantics can fulfill this role by providing a measure of semantic similarity between words that is derived from lexical co-occurrences in large text corpora. On the basis of a case study of the construction "V *the hell out of* NP", e.g., *You scared the hell out of me*, it is shown that distributional semantics not only appropriately captures how the verbs in the distribution of the 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 determinants of syntactic productivity in naturally occurring data.

Keywords: syntactic productivity, constructions, distributional semantics, vector-space model, diachrony, language change, usage-based

1 Introduction

Grammars are in constant change. Over time, different ways are created in which words can be combined into phrases and sentences, while others fall into disuse. For example, in English, the basic word order used to be SOV at the onset of the Old English period, during which the language went through a gradual shift in word order (Hock and Joseph 1996: 203–208). The SVO order found nowadays initially emerged from an innovation involving the displacement of auxiliaries to second position for prosodic reasons (cf. Dewey 2006), which was later re-analyzed as concerning all finite verbs. The older SOV order persisted for some time, notably in dependent clauses, but had almost completely disappeared by the end of the Middle English period.

***Corresponding author: Florent Perek,** Department of English, University of Basel, Nadelberg 6, CH-4051 Basel, Switzerland, E-mail: florent.perek@gmail.com

Beside such drastic and long-lasting shifts in 'core' aspects of grammar, language change may also consist of more subtle variation in usage. As reported by many studies, in the course of no more than a few decades, speakers of the same language might show slightly different preferences for the grammatical means they use to convey the same message (cf. Aarts et al. 2013; Krug 2000; Leech et al. 2009; Mair 2006). For example, Mair (2002) finds that the bare infinitive complementation of *help* (e.g., *Sarah helped us edit the script*) has increased in frequency between the 1960s and the 1990s in both British and American English, compared to the equivalent *to*-infinitive variant (e.g., *Sarah helped us to edit the script*). Observations of this kind are often regarded as grammatical change in the making.

Among the facts about usage that are subject to diachronic change, this paper is concerned in particular 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 (cf. Goldberg 1995), exemplified by (1) and (2) below (see also Traugott and Trousdale 2013: 86–91).

(1) *They hacked their way through the jungle.*

(2) *She typed her way to a promotion.*

In both examples, the construction conveys the motion of the agent, which is metaphorical in (2), along the path described by the prepositional phrase, and the main verb conveys the means whereby motion is enabled. As Israel points out, this use of the construction is attested as early as the sixteenth century, but it was initially limited to verbs describing physical actions (e.g., *cut* and *pave*), with which the construction conveys the actual creation of a path enabling motion, like in (1). It was not until the nineteenth century that examples like (2) started to appear, in which the action depicted by the verb provides a markedly more indirect way of attaining the agent's goal. Similar cases cited by Israel (1996: 224) involve the verbs *write*, *spell*, and *smirk*.

This paper presents an attempt to study syntactic productivity in diachrony in a fully data-driven way. As reported in Section 2, most contemporary approaches to syntactic productivity emphasize the role of semantics, which poses the methodological problem of defining and operationalizing word meaning and semantic similarity. As discussed in Section 3, none of the current solutions is entirely satisfactory. In Section 4, an alternative solution to this

methodological problem is described that makes use of distributional information as a proxy to meaning in order to derive a measure of semantic similarity. In Section 5, this method is applied to a case study of the construction "V *the hell out of* NP" (e.g. *You scared the hell out of me*) in American English. It is shown that a distributional approach to word meaning not only is adequate for the study of syntactic productivity, but also presents the advantage of allowing the use of visualization techniques and statistical analysis.

2 Determinants of syntactic productivity

The notion of productivity has a long history in the field of morphology, where it refers to the property of a word formation process to be used by speakers to coin new words. For example, the suffix *-th*, as in *length*, *health*, and *growth*, cannot be used in modern English to form new nominalizations, and is therefore to be considered as not (or no longer) productive, whereas the suffix *-ness* is readily available to speakers for deriving a noun from an adjective (cf. Plag 2003: 44–45). A prime example would be *nouniness*, describing the extent to which a word behaves as a noun, which was, to my knowledge, first coined by Ross (1973) from the adjective *nouny*, itself productively derived from the adjective-forming suffix *-y* and the noun *noun*.

It is only in recent years that a similar notion was applied to the domain of syntax, which had long been dominated by the conception of grammar as a system of abstract rules separated from the lexicon that enables speakers to produce an infinite number of sentences, including those that they have neither used nor heard before (cf. e.g., Chomsky 1986). Under this view, lexical items can be freely combined with syntactic structures as long as the former match the grammatical specifications of the latter. However, studies of language use have made it increasingly clear that words and syntactic constructions combine in non-trivial ways, in that the slots of constructions are not equally likely to be filled by any lexical item, even when the resulting combination would make perfect sense. It is often the case that combinations that could be expected either do not occur (or marginally so), or are even judged unacceptable. For example, Goldberg (1995: 79) notes that the adjective slot ("Adj") in the construction [*drive* NP Adj], e.g., *The kids drove us crazy*, is mostly restricted to describe a state of insanity, such as *crazy*, *mad*, *insane*, *nuts*, etc. Hence, even though the construction itself conveys a resultative meaning (i.e., 'X causes Y to become Z') which could in principle combine with any predication in a sensible way, most instances with other kinds of adjectives (especially positive ones), like **drive*

someone successful, are unacceptable. Hence, in much the same way as morphological patterns, syntactic constructions, or rather, their "slots", display varying degrees of productivity. Zeldes (2012) reports similar findings in the domain of argument selection, in that semantically similar verbs may be very different in their tendency to be used with novel arguments. Under this view, fully productive, unconstrained constructions, as per Chomsky's (1986) definition of productivity, occupy one extreme end of a continuum of productivity, with more restricted constructions such as [*drive* NP Adj] lower down the scale, and fairly lexically-specific patterns at the other extreme (e.g., *look/stare someone in the eye* but not **gaze/peer someone in the eye(s)*).

The notion of syntactic productivity can be applied to the description of various kinds of linguistic behavior. In language acquisition, it accounts for the children's ability to generalize beyond their inherently limited input by producing combinations of words and constructions they might not have witnessed in the speech of their caregivers (Bowerman 1988; Pinker 1989). Of course, children are not the only productive speakers, as adults too do at times use the resources of their language creatively to produce unconventional expressions (cf. Pinker 1989 for some examples). By the same token, the concept of syntactic productivity also applies in synchrony when new words enter the language, in that there may be several constructions in competition for using a novel word in a sentence (cf. Barðdal 2008 for examples with recent Icelandic verbs related to information technology and to novel means of communication). Finally, in language change, syntactic productivity may refer to the property of the slots of a syntactic construction to attract new lexical fillers over time, thereby forming novel combinations of words and constructions (Barðdal 2008); it is this last aspect that will be the focus of this paper. While it is not clear whether these various phenomena should be considered to involve the same underlying process, there is evidence that at least some of them are driven by similar factors (cf. Barðdal 2008; Suttle and Goldberg 2011; Wonnacott et al. 2012; Zeschel 2012), although there might be differences in the relative importance of these factors.

At first blush, syntactic productivity appears to be partly arbitrary, in that, as was just pointed out, there can be many combinations of lexemes and constructions that make perfect sense following compositional semantics but are nevertheless never uttered, if they are considered acceptable at all. However, a growing body of evidence seems to indicate that the productivity of a construction is ultimately tied to the previous experience of speakers with that construction. In this usage-based view, it has been proposed very early on that syntactic productivity is promoted by high type frequency, i.e., by a high number of different items attested in the relevant slot of a construction (Bybee and Thompson 1997; Goldberg 1995). This hypothesis is motivated by findings

from morphology (Bybee 1985, Bybee 1995), thus drawing a parallel between the two domains. The idea makes intuitive sense: speakers should be more confident that a pattern can be extended to new items if they have witnessed this pattern with many items than if they have seen it restricted to only a few. However, if this intuition is correct, it is clear that the diversity of items matters at least as much as their sheer number, as pointed out by Goldberg (2006). Since an increase in type frequency is usually correlated with an increase in variability, type frequency provides an indication of a pattern's degree of productivity, but is not necessarily the source of this productivity. Under this view, a pattern is only productive to the extent that it instantiates a high number of dissimilar items.

Barðdal (2008) combines the two factors (type frequency and semantic variability) by proposing that productivity is a function of the inverse correlation between type frequency and semantic coherence (i.e., the inverse of variability), in that the relevance of type frequency for productivity decreases with semantic coherence. Hence, a construction witnessed with few items will only be productive if these items are highly similar, and, even so, will only allow novel uses within the restricted semantic domain defined by its distribution; the construction [*drive* NP Adj] mentioned above falls into this category (cf. Bybee 2010). Conversely, a construction occurring with highly dissimilar items will not necessarily allow novel uses, in that the semantic variability must be attested by a sufficient number of items (high type frequency). These two types of constructions, i.e., low type frequency and high semantic coherence vs. high type frequency and low semantic coherence, correspond to two kinds of productivity at the extreme ends of the continuum, that are traditionally kept apart in the literature, respectively analogy-based productivity and schema-based (or rule-based) productivity, which Barðdal sees as two sides of the same coin.

In line with this usage-based approach to productivity, many studies report that the occurrence of a novel item in a construction seems to depend on its similarity to previous usage. Barðdal (2008: Ch. 3) finds that novel verbs related to information technology in Icelandic come to be used in an argument structure construction with which semantically similar verbs are already attested. Similarly, Bybee and Eddington (2006) find semantic similarity with frequent attested instances to be a significant determinant of the acceptability of infrequent uses of Spanish Verb-Adjective copular constructions with the verbs of becoming *quedarse* and *ponerse*. They take their findings as providing evidence for an exemplar-based model of grammatical constructions in which productivity is driven by analogy with previously stored instances (see also Bybee 2010). Barðdal (2008) also draws on analogical extensions, on the basis of individual verbs or lower-level constructions corresponding to clusters of semantically

similar verbs, to explain the occasional occurrence of particular novel verbs in constructions with low type frequency when a more type-frequent alternative exists. On the other hand, she also reports cases where, even in the presence of a highly similar item in the distribution of a construction, some novel verbs are used in another construction with high type frequency and semantic variability, which seems to act as a kind of highly productive 'default' construction, recruited more or less independently of lexical semantics (cf. Barðdal 2008: Ch. 4, Barðdal 2011). This indicates that there does not always have to be a similar item in the distribution of a construction for a new coinage to occur, in that a construction can become virtually open to any compatible item once it has been attested with a sufficient amount of variability.

In an experimental study, Suttle and Goldberg (2011) aim to tease apart type frequency, variability, and similarity, and evaluate the respective contribution of these three factors to syntactic productivity. Participants were presented with sets of sentences in a fictitious language. All sentences consisted of two noun phrases with a definite article and a novel nonsense noun, and an English verb followed by a nonsense particle (e.g. *The zask the nop toast-pe*). Each set of sentences exemplified a different construction; the constructions differed according to the position of the verb (initial, medial, or final) and the particle, which was unique for each construction. After each exposure set, participants were asked to rate the likelihood of another instance of the same construction with a different verb (the target). With this design, Suttle and Goldberg could systematically manipulate the three following factors: (i) type frequency, by varying the number of different verbs in each set, (ii) variability, by choosing the stimuli verbs from the same vs. three different semantic classes, as determined by Levin's (1993) classification, and (iii) similarity, by choosing the target verb from one of various semantic classes, respectively one represented in the stimuli set (high similarity), vs. a similar class, i.e., concrete actions for both stimuli and target verbs (moderate similarity), vs. an unrelated class, i.e., verbs of cognition for the target (low similarity). They found that type frequency and variability each have an independent effect, and that they are also involved in a positive interaction, i.e., the effect of type frequency is stronger when variability is high. This finding is in line with Barðdal's idea that type frequency is a more important predictor of productivity for highly variable constructions than for semantically coherent constructions. They also report a main effect of similarity, in that the closer a coinage is to an attested utterance, the more acceptable it is found by participants. Interestingly, the effect of variability also varies greatly according to the degree of similarity of the target verb to the stimuli verbs. When similarity is high, the effect of variability is negative, in that subjects are more confident about the coinage when the distribution of the construction is less variable, which Suttle

and Goldberg suggest is because participants in the low variability condition see more verbs from the same class attested in the construction (since type frequency was held constant), and therefore receive more evidence suggesting that any verb from this class may be used in the construction. With moderate similarity, there is a positive effect of variability, showing that the acceptability of unattested classes improves when there is evidence that the construction is already attested in multiple, relatively similar classes. However, when similarity is low, there is no effect of variability, which means that variability is irrelevant when the target bears no resemblance to any attested item.

To explain their results (especially the complex interaction between variability and similarity), Suttle and Goldberg (2011) propose the notion of coverage that they define as "the degree to which attested instances 'cover' the category determined jointly by attested instances together with the target coinage" (Suttle and Goldberg 2011: 1254). This notion is reminiscent of Clausner and Croft's (1997: 263) definition of the degree of productivity of a schema as "[t]he proportion of [its] potential range [i.e., the range of concepts consistent with the schema] which is actually manifested". In addition, the concept of coverage also calls attention to how the semantic domain of a construction is populated in the vicinity of a given target coinage, and more specifically to the density of the semantic space. If the semantic space around the novel item is dense (high coverage), 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 (lower coverage), the less likely this item can be used, which is in no small part related to variability, since if the same number of items are more "spread out" around a given coinage, the density of the semantic space will decrease. Hence, Suttle and Goldberg's proposal conceives of productivity not as an absolute property of a construction, but as a phenomenon that takes into account the relation between attested items and potential instances. Following the notion of coverage, a construction can rarely be said (if ever) to be productive in absolute terms; rather, a construction is productive to various degrees in different semantic domains.

To summarize, the view of productivity that emerges from previous research is that of a phenomenon that is strongly tied to the previous usage of constructions. In a nutshell, speakers are likely to use a construction in similar ways to its priorly witnessed usage, unless the structure of its distribution invites them to depart from it. Importantly, previous studies point to a strong semantic component, in that novel uses must be semantically coherent with prior usage. 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 gives rise to a number of methodological issues, described in the next section.

3 Factoring in semantics in studies of syntactic productivity

As previously mentioned, most current accounts of syntactic productivity heavily rely on semantics. Consequently, any attempt to test these models against empirical data requires an operationalization of the semantic aspects of productivity. More specifically, it requires an assessment of such aspects of meaning as variability in a set of items and similarity between items. This section discusses how meaning can be captured in empirical research on productivity, and describes the methodological and practical issues involved (see also Zeschel 2012 for a similar discussion). An alternative is suggested that relies on distributional semantics.

The motivation behind the present research stems from a simple observation: linguistic meaning is not directly observable in the same way that morphosyntactic or phonetic properties are. This is especially true for corpus studies: a corpus only ever contains signifiers (or forms), and accessing the meaning of these forms requires a human interpreter. Since searching a corpus, as opposed to, for instance, collecting behavioral data from native speakers, is the only way to observe earlier stages of a language, the issue of factoring in semantics is inescapable for the study of syntactic productivity from a diachronic perspective, which is the one adopted in this paper.

The most basic and probably the most common way to deal with meaning in corpus studies is for the analyst to perform manual semantic annotation. This can take a number of forms, depending on the requirements of the study: from adding simple semantic features, such as animacy or concreteness, to more subtle judgments such as word sense distinctions, and even more complex tasks like identifying instances of an abstract construction. Some semantic annotation tasks can be facilitated by deriving the annotation scheme from an external source, such as Levin's (1993) verb classes or the lexicographic database WordNet (Miller 1995), although the efficiency of such sources may be limited by their coverage.

Importantly, manual annotation primarily produces categorical data, as the judgments it is based on consist in deciding which category a given item belongs to. As such, it does not allow to directly derive gradient measures of similarity and variability, which limits its usefulness for studies of syntactic productivity. Surely, it is in principle possible for the analyst alone to estimate degrees of similarity between items or posit semantic groupings, but such data are not clearly amenable to precise quantification and hardly reflect the complexity of the semantic space. It should be acknowledged that a form of similarity measure can be derived from categorical data if the annotation scheme specifies relations

between categories. This is the case of the WordNet database, which encodes various kinds of relations between word senses, such as hyperonymy/hyponymy, entailment, meronymy, etc. The graph structure of WordNet, and in particular the taxonomic hierarchy it defines through hyperonymy/hyponymy relations, can be used to derive various kinds of similarity measures based on the number of graph edges that need to be traversed to connect two word meanings (cf. Budanitsky and Hirst 2006). These WordNet-based similarity measures are conceptually not unproblematic, since the taxonomic hierarchy primarily reflects the structure of the lexicon, but not necessarily the semantics of the words it contains. The hyponymy relation is given the same weight regardless of the amount of semantic information that separates the words it connects, which may vary across such word pairs. Moreover, the precision of semantic distinctions is limited by the range of lexical concepts that receive a distinct word form in the language, and in particular by the most abstract level that is still lexicalized in a given hierarchy, above which everything is merged meaninglessly into an empty "root" node, with no possible way to retrieve intermediate degrees of similarity. Despite these inherent limitations, WordNet-based similarity measures seem to achieve good performance at least for nouns, but they have been much less systematically tested on other parts of speech. The usability of WordNet for studies of syntactic productivity is not explored in this paper, but it is certainly a topic worthy of further investigation.

More generally, manual semantic annotation poses the methodological problem that it is based on the semantic intuitions of a single individual, which renders it potentially subjective: different annotators might disagree as to how to categorize items, or what pairs of items are more similar to each other. Admittedly, the issue can be addressed by assigning the task to several annotators and checking for agreement. Along these lines, Bybee and Eddington (2006) conducted a semantic norming study in which similarity judgments were collected from a group of native speakers. Participants were presented with pairs of items and asked to rate how similar they found these items on a given scale. Zeschel (2012) suggests a refined version of this task that takes into account various kinds of semantic relations, such as antonymy, hyponymy, and metonymical shift. By pooling the data from all participants/annotators, a more objective (or at least intersubjective) measure of semantic similarity can be obtained that should be more faithful to the intuitions of the linguistic community. What is more, this measure lends itself directly to quantitative analysis. The norming study design is probably the soundest way to assess semantic similarity, both theoretically and methodologically, but it is decidedly less convenient and probably more time-consuming than manual annotation, not to mention that it necessitates access to a population of native speakers ready to provide

semantic judgments, possibly for a compensation.[1] More importantly, it is also inherently limited in scope in that it is constrained by the number of judgments that one may reasonably collect from a single speaker. Since each item from the set under consideration must be compared to every other item, the number of judgments grows exponentially with the size of the set and quickly reaches a number that makes the study practically unfeasible. Bybee and Eddington sidestep this issue by limiting their study to 20 items, which already requires 190 judgments. By way of comparison, 50 and 100 items, which even a moderately productive construction easily attains, respectively require 1,225 and 4,950 judgments, ideally per participant. In sum, while a norming study is the most appropriate solution in theory, it is in practice not applicable to a great many examples of constructions.

In the light of these issues, this paper evaluates another possible solution to the problem of assessing semantic similarity that was already mentioned by some scholars but has not yet, to my knowledge, been explored further for the purpose of studying syntactic productivity. Throughout its history, it has been common for corpus linguistics to borrow various techniques from neighboring fields to handle corpus data, especially from computational linguistics. It is especially true in the case of automatic annotation, which is nowadays commonly used to add additional layers of linguistic information to electronic corpora, such as part of speech, lemma, or syntactic structure. To the extent that it fulfills its purpose with enough accuracy, automatic annotation eschews the need for manual checking by human annotators, which is costly and time-consuming. Along similar lines, the present study shows how distributional semantics and its main computational implementation, the vector-space model, can also be fruitfully used to augment corpus data with information about lexical meaning in an automatic, data-driven way that, to a large extent, dispenses with the need for human semantic intuitions. This technique is described in the next section.

1 Note that this requirement has been relaxed by the advent of online experiments (cf. the WebExp system, http://www.webexp.info/ [consulted Feb 7 2014]), which are gaining increasing acceptance as appropriate sources of empirical data in psychology. The World Wide Web provides researchers with a wealth of participants for their studies, and, importantly, dispenses with considerations of time (any number of subjects can participate at the same time and at any moment) and space (anybody in the world with an Internet connection can participate). In particular, Amazon Mechanical Turk provides a platform both for posting online experiments and surveys and for recruiting subjects that is growing increasingly popular among psychologists.

4 Distributional semantics and vector-space models

Distributional semantics is the dominant and to this day most successful approach to semantics in computational linguistics (cf. Lenci 2008 for an introduction). It draws on the observation that words occurring in similar contexts tend to have related meanings, as epitomized by Firth's (1957: 11) famous statement "[y]ou shall know a word by the company it keeps". Therefore, a way to access the meaning of words is through their distribution (cf. Miller and Charles 1991 for experimental evidence supporting this view). For example, the semantic similarity between the verbs *drink* and *sip* will be seen in their co-occurrence with similar sets of words, such as names for beverages (*water*, *wine*, *coffee*), containers (*glass*, *cup*), or, more subtly, words related to liquids (*pour*, *hot*, *cold*, *steaming*) and dining/drinking practices (*table*, *chair*, *bar*, *counter*, *dinner*, *restaurant*). This is not to say that *drink* and *sip* will not share some of these collocates with other, more distantly related words (like for instance *spill*), but because *drink* and *sip* are so similar, it is expected that their distribution will show a particularly high degree of overlap in a corpus of sufficient size. In sum, in distributional semantics, the semantic similarity between two words is related to the number of their shared frequent collocates in a vast corpus of naturally occurring texts.[2] Conversely, differences in the distributional profile of two words are expected to correspond to differences in their meaning.

Vector-space models are the main technical implementation of distributional semantics (Turney and Pantel 2010; Erk 2012). They owe their name to the fact that they derive semantic information by associating words with arrays of numerical values (i.e., vectors) based on co-occurrence counts. The first step in creating a vector-space model is to build a co-occurrence matrix, with the set of

2 According to Sahlgren (2008), this conception of distributional semantics captures paradigmatic similarity in particular, i.e., the extent to which words can be substituted in context, as opposed to syntagmatic similarity, i.e., the extent to which words tend to co-occur in the same units of text. The latter kind of similarity is captured by vector-space models that take the frequency of occurrence of words in documents as input; hence, each column corresponds to one document, and words occurring in the same documents are judged more similar. An example of document-based vector-space semantic modeling is Latent Semantic Analysis (Landauer et al. 1998). As it turns out, syntagmatic similarity tends to relate words involving similar topics (e.g., *hospital*, *doctor*, *nurse*), and semantically similar verbs are rarely related in this way. Hence, paradigmatic similarity is more appropriate for the case study presented in this paper, and, more generally, better captures the kind of semantic relatedness that is relevant to syntactic productivity.

words under consideration as rows, and the collocates against which the meaning of these words is assessed as columns. The matrix is filled by counting, for each occurrence of the target words in a corpus, their frequency of co-occurrence with other words within a set context window. Function words (articles, pronouns, conjunctions, auxiliaries, etc.) and other semantically near-empty items, such as numbers or frequent modifiers (*very*, *really*), are usually ignored, as they are assumed not to contribute to the identification of relevant semantic distinctions, and would therefore only be a source of noise if they were included. A frequency threshold is also often used to avoid data sparsity. For example, Table 1 below presents a co-occurrence matrix for *drink* and *sip* based on the mini-corpus given in Figure 1, which contains three occurrences of these two verbs in the Corpus of Contemporary American English (Davies 2008) in a five-word context window (i.e., five words to the left and five words to the right).

Such a small sample is obvious not enough to make any robust claims about the meaning of *sip* and *drink* on the basis of their distribution, but some basic trends are already visible.[3] As expected, both words co-occur with names for beverages: *beer*, *champagne*, *water*; other words related to drinking and dining practices are found: *food*, *glass* (two words also related to beverages), *pizzeria*, *table*. The two verbs share three of these collocates: *beer*, *cold*, and *glass*; with a larger sample, we would probably obtain more shared collocates of the same kind, while the other cells would remain mostly empty. This example illustrates the idea that the distribution of words reflects aspects of their meaning.

Various kinds of transformations are usually applied to the co-occurrence matrix. Weighting employs information-theoretic measures, such as point-wise mutual information, to turn 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, i.e., to what extent the collocate occurs with that word more often than with other words. Also, dimensionality reduction can be employed to transform the matrix so that it contains fewer columns, selecting and consolidating the most salient contextual features by means of linear algebra such as singular value decomposition. In addition to making operations on the matrix computationally more tractable, dimensionality reduction also singles out the most informative aspects of word distributions.

In the (possibly transformed) co-occurrence matrix, each row is a word vector, which represents the distributional profile of this word. Under the

3 Admittedly, these contexts were carefully selected for the sake of the example, but it would not be hard to reproduce the same trends on randomly selected instances, although a much larger number would be necessary.

Table 1: Co-occurrence matrix for the verbs *drink* and *sip* based on the mini-corpus given in Figure 1.

	beer	book	champagne	change	cold	dress	food	glass	hell	intuition	man	meet	pick	pizzeria	put	table	trust	water	while
Drink	2	1	0	1	1	1	0	2	1	0	0	1	0	1	0	1	0	1	1
Sip	2	0	1	0	1	0	1	1	0	1	1	0	1	0	1	0	1	0	0

```
    the pizzeria for a while,  drinking  a beer at a table
          hell, I'd meet you,  drink     a glass of beer or
books. She changed her dress,  drank     a glass of cold water

   men picked up their beers,  sipped    them, and put them back
  to trust his intuition. She  sipped    from the champagne glass and
    food itself. Even when he  sipped    his cold beer, it was
```

Figure 1: Three occurrences of *drink* and *sip* from the COCA.

assumption that semantic distance between words is a function of distributional differences, similarity between rows approximate semantic similarity, which can be quantified by mathematical measures. In that connection, the co-occurrence matrix is often conceptualized as representing a multi-dimensional semantic space, in which each word receives coordinates according to its distribution. To derive semantic similarity, or its converse, semantic distance, the cosine measure is by far the most frequently used in distributional models of word meaning.[4] Its main advantage is that it normalizes for word frequency, in that two words from a different frequency range will be judged similar if their collocates occur with proportionally similar frequencies, even though the raw frequencies of co-occurrence might differ substantially.

A caveat should be added at this point. The term "semantic similarity" might not be the most fitting to describe the measure derived from distributional information, as it should not be interpreted as entailing synonymy. Indeed, groups of words that are found most similar according to distributional semantics are not necessarily synonyms. Antonyms, for instance, are often found to be similar in distributional models precisely because they tend to co-occur with the same words, as a reflection of the semantic component that they share, i.e., the scale on which they are opposites. Technically, distributional similarity reflects the extent to which two words can be substituted for each other, which might capture different aspects of their meaning. Besides synonymy and antonymy, other kinds of semantic relations can cause words to occur in similar contexts, such as co-hyponymy and hyperonymy. In sum, the semantic measures derived from distributional information should be considered measures of unspecified semantic relatedness rather than semantic similarity proper. This does not, however, undermine the usability of this measure in the context of syntactic productivity, since various kinds of semantic relations have been found to matter for this phenomenon (cf. Zeschel 2012).

A common criticism leveled at vector-space models is that they ignore polysemy, in that distributional information is assigned to word forms, and thus each word form is associated with a single semantic representation. While this comment is in order, whether or not it is an actual problem for a particular application is an empirical question. The problem does obviously not arise with monosemous

4 In mathematics, the cosine function varies between 0 and 1 in positive space (like the distributional space derived from frequency counts). A cosine of 1 means that the vectors are identical, in the sense that they point to the same direction; a cosine of 0 means that they are orthogonal, i.e., maximally divergent. Therefore, the cosine is *per se* a measure of similarity. To turn it into a distance measure, as required by the analyses reported in Sections 5.4, 5.5, and 5.6, the cosine values can be subtracted from 1, i.e., $distance_{\cos}(V_1, V_2) = 1 - cosine(V_1, V_2)$.

words, and it is often not problematic to consider related or similar senses as a single meaning; the issue is of course more serious in the case of true homonymy. It is also not uncommon that the distribution of words with multiple senses is dominated by a single sense in corpora. In that case, polysemy can be seen as a mere source of noise for the assessment of that particular sense. Truly polysemous words, i.e., with clearly differentiated senses balanced in frequency, should be treated with a grain of salt, since they will tend to be considered mildly similar to several different words. Some researchers have suggested methods to identify multiple word senses in distributional information (Pantel and Lin 2002; Purandare and Pedersen 2004; Schütze 1998). In this study, the polysemy issue can largely be ignored, since most of the verbs submitted to distributional classification have a low degree of polysemy.

The main benefit of vector-space models over other, non-quantitative approaches to word meaning is that the informal notion of semantic representation is turned into an empirically testable semantic model. In such an approach, semantic similarity can be quantified, which opens a range of useful applications for empirical studies, such as the derivation of other quantitative measures or statistical testing (cf. Section 5.6). Also, while the status of distributional semantics as a theory of semantic representation and acquisition is still much debated (cf. Glenberg and Robertson 2000), distributional models have been argued to display some potential for psychological reality. Some implementations have been shown to correlate positively with human performance on various tasks, such as synonymy judgments, word association, and semantic priming (Lund et al. 1995; Landauer et al. 1998), which means that they are at least good models of human behavior. Andrews et al. (2008) evaluate the relative importance of distributional knowledge and experiential information (i.e., based on properties available to the senses) for semantic representations, by comparing the performance of models based on each kind of information with one based on a combination of both. They find that a model based on both kinds of information provides more coherent results and also performs better on a set of comparisons with human-based measures of semantic representation (lexical substitution errors, association norms, semantic priming in word recognition, and interference in word production). These results suggest that distributional information might well be a key component of how human beings acquire and process semantic information. Hence, the present study's attempt to use a distributionally derived measure of semantic similarity to study syntactic productivity does not only address the practical concern of obtaining semantic information without relying on human intuitions: it might also qualify, to some extent, as a cognitively grounded approach to the issue. That being said, it should be emphasized that vector-space modeling is merely seen as providing a proxy to word meaning in

this paper, which remains agnostic as to whether distributional information should be considered as a cognitive representation of meaning itself.

Vector-space models are widely known in computational linguistics and have been used for many practical applications, including word-sense disambiguation (Pedersen 2006), automatic thesaurus generation (Grefenstette 1994), and information extraction (Vyas and Pantel 2009). Yet, while distributional information of any kind is used increasingly commonly by linguists to ground linguistic generalizations in patterns of usage (e.g., Divjak and Gries 2006; Croft 2010; Wälchli and Cysouw 2012), distributional semantics in particular has been much less frequently employed in theoretically-oriented work. Among the rare occurrences, Gries and Stefanowitsch (2010) draw on distributional semantics to inductively identify verb classes in the distribution of constructions by clustering verbs according to their frequent collocates. Similarly, Levshina and Heylen (2014) use a vector-space semantic model to identify contrasting sets of semantic classes for the causee argument in Dutch periphrastic causative constructions with *doen* and *latten*. In historical linguistics, distributional semantics has been used by some scholars to track recent semantic change (Boussidan 2013; Cook and Stevenson 2010; Gulordava and Baroni 2011; Sagi et al. 2009). However, no attempt has yet been made to apply distributional semantics to the study of syntactic productivity in diachrony. This paper seeks to mend this gap. As will be shown, adopting distributional methods to the problem of handling semantic information is an empirically appropriate solution to the issues mentioned in the last section. As a result, it increases the scope of possible studies, since it raises the constraint on the number of lexemes that can be considered. The next section presents a case study demonstrating the appropriateness of a distributional approach to lexical semantics for the study of syntactic productivity, and the analytical advantages that it offers.

5 Case study

5.1 The *hell*-construction

The case study presented in this paper considers the construction corresponding to the syntactic pattern "V *the hell out of* NP" (Hoeksema and Napoli 2008; Haïk 2012), as exemplified by the following sentences from the Corpus of Contemporary American English (hereafter 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) *Damn this man loved the hell out of his woman.*

(6) *Me and Jeff want to beat the hell out of each other.*

(7) *You might kick the hell out of me like you did that doctor.*

The construction is typically used with two-participant verbs, and basically consists in a two-argument construction where the post-verbal argument is preceded by the phrase *the hell out of*. Compared to a regular transitive construction, the *hell*-construction generally conveys an intensifying function, very broadly defined. The examples above illustrate the most common and straightforward case, in which the construction intensifies the effect of the action or the effort of the agent. Hence, *scare/surprise/love the hell out of* means "scare/surprise/love very much", and *beat/kick the hell out of* means "beat/kick very hard". Examples (8) and (9) below with *play* and *wear* exemplify another less common, though relatively regular case, in which the action is a performance, or is construed as such in the case of *wear*, and it is the quality of this performance that is intensified.[5]

(8) *Phil and his music-mates [...] could play the hell out of any song.*

(9) *[A]wards-show-bound actors and directors show twelve different ways to wear the hell out of a tuxedo.*

In some cases, the particular aspect that is intensified may be highly specific to the verb and may depend to some extent on the context. With *ride* in (10), the event is longer and involves more strain on the vehicle than usual. In (11), *sell the hell out of* means (in this case) "sell a lot". Both examples relate to the intensification of the agent's effort, as mentioned previously, and so do examples (12) and (13), which also focus on the insistence of the agent to obtain a particular result.

(10) *Our test team rode the hell out of these bikes on your behalf.*

(11) *By then I was selling the hell out of Buicks at Petz's.*

5 Examples (8) to (13) are also from the COCA.

(12) *I kept Googling the hell out of 'stress fracture' and 'femoral neck'.*

(13) *If you ever hear that I've committed suicide, investigate the hell out of it.*

Instances of the *hell*-construction superficially look as if they consist of a direct object followed by a prepositional phrase headed by the complex preposition *out of*, and therefore appear to be instances of the removal construction (Goldberg 2011) conveying the meaning 'X CAUSES Y to MOVE from Z', e.g., *He took the gun out of the holster*. In fact, as argued by Hoeksema and Napoli (2008), the *hell*-construction probably emerged from uses of the removal construction describing an exorcism scenario, specifically expressions like *beat the devil out of X*, literally meaning 'causing a demon to leave a person's body by beating them'. At some point towards the end of the nineteenth century, expressions of this kind started to undergo semantic bleaching as they were used in contexts where they did not convey an exorcism scenario literally but only through hyperbolic reference, for the purpose of intensification, as in (14) below.

(14) *Yes, Loubitza will beat the devil out of her when she gets her home – her and her broken jar!* (1885; cited by Hoeksema and Napoli 2008: 371)

Concomitantly, the phrase *the devil* in this and other constructions was progressively being replaced, by metonymy, with *the hell*, and this new expression gave rise to the modern *hell*-construction, which retained the bleached meaning of intensification, and became increasingly popular from the 1930s onwards (cf. Section 5.2).

While the *hell*-construction probably arose from a literal meaning of removal, there is evidence that it developed into a construction that no longer patterns semantically and syntactically like instances of the removal construction. First, as noted by Haïk (2012), the preposition in the *hell*-construction is restricted to be *out of*, and other prepositions with a similar 'removal' meaning like *off* and *from* are barred from it, as shown in (15a). Such restrictions do not hold for the removal construction, as shown in (15b).

(15) a. **He kicked the hell off/from me.*
b. *He took the gun off/out of/from her hands.*

Secondly, while there seems to be evidence that the phrase *the hell* has some syntactic properties of direct objects (in particular with regards to passivization, cf. Hoeksema and Napoli 2008; Haïk 2012), its semantic status as a theme argument in the removal construction is highly questionable. Evidently,

sentences (3) to (13) do not involve caused motion, and, in fact, *the hell* is not even referential (even in a figurative sense), as shown by the impossibility of anaphoric reference by a pronoun (cf. [16]).

(16) **He scared the hell out of Sam, and kicked it out of Bill too.*

This questions whether *the hell* should be treated as a noun phrase in the traditional sense. It rather appears to be an instance of some kind of expressive phrase that can also be found in other expressions to convey a similar exclamatory function, e.g., *What the hell is going on?*, *get the hell out of here*, *for the hell of it* (cf. also expressions like *one hell of a mess*). Added to the fact that the referent of the prepositional phrase complement bears the same semantic role as the direct object of the transitive counterpart, this suggests that the *hell*-construction could be treated as a case of particle insertion, or as an alternation à la Levin (1993), whereby the expression *the hell out of* is inserted before the direct object argument and modifies the predication in a quasi-compositional way. However, while the overwhelming majority of verbs occurring in the construction are transitive, uses with intransitive verbs are also attested, such as *listen* in (17a). *Listen* cannot be used transitively as in (17b); its non-subject argument must be preceded by the preposition *to*, as in (17c). Hence, there is no transitive construction from which example (17a) could be derived through the addition of the phrase *the hell out of*.

(17) a. *I've been listening the hell out of your tape.* (COCA)
b. **I've been listening your tape.*
c. *I've been listening to your tape.*

Taken together, these observations suggest that the pattern cannot be derived compositionally from any other constructions in the language, and therefore forms its own generalization. As also noted by Hoeksema and Napoli (2008: 375), the *hell*-construction lends itself nicely to a construction grammar analysis (Goldberg 1995), whereby the abstract meaning of intensification is directly associated with the whole phrasal pattern "V *the hell out of* NP". As such, the *hell*-construction can be seen as a case of constructionalization, in which an instance of an existing construction gradually evolved into an independent generalization (Bybee 2013; Traugott and Trousdale 2013). The *hell*-construction is similar to more "vulgar" variants in which *hell* is replaced by other taboo words (e.g., *crap*, *fuck*, *shit*; see Hoeksema and Napoli 2008 for a thorough list of attestations), and could therefore be considered a member of a family of related constructions (Goldberg and Jackendoff 2004).

5.2 Corpus data

This study uses the Corpus of Historical American English (COHA; Davies 2010) as a source of data on the diachronic development of the verb slot in the *hell*-construction. The COHA consists of about 20 million words of written American English for each decade between 1810 and 2009 and is available online.[6] The corpus is roughly balanced for genre, in that each decade contains texts of four types (fiction, magazines, newspapers, non-fiction) in about the same proportions.[7] The string "[v*] the hell out of" was searched for in the COHA, which returned instances of all verbs followed by the sequence "the hell out of". All tokens were downloaded and the instances of the *hell*-construction were filtered out manually, 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 several diagrams in Figure 2. Most of the first attestations of the construction in the corpus

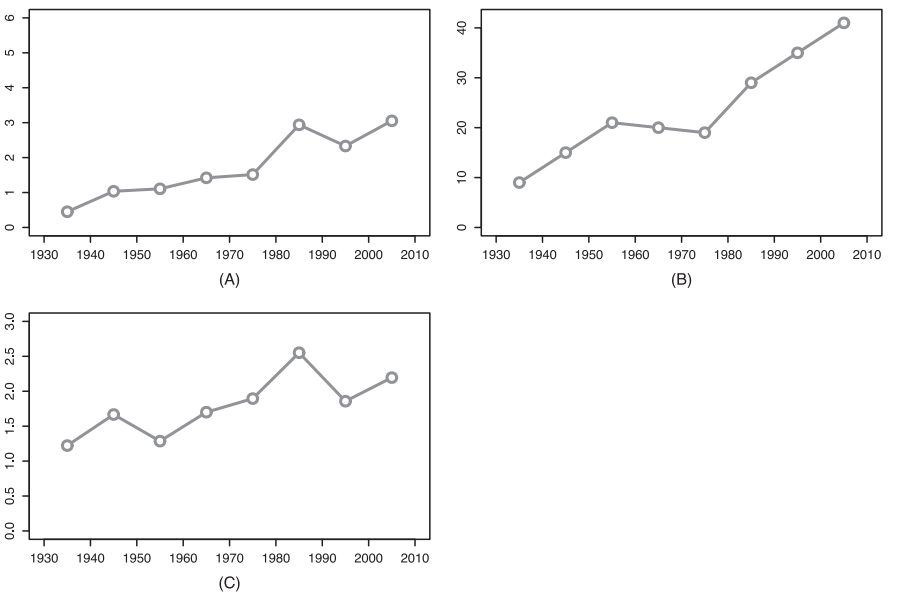

Figure 2: Diachronic development of the *hell*-construction in token frequency normalized by million words (A), type frequency (B), and token/type ratio (C), per decade.

6 Twenty million words is a rough average; recent decades tend to be markedly bigger (there are no less that 29 million words for the 2000s), and the earliest sections smaller.

7 This is true at least for all decades from the 1870s onwards; before that, the corpus contains little to no newspaper data, and the other genres are balanced slightly differently. See http://corpus.byu.edu/coha/help/texts_e.asp (consulted Feb 7 2014) for details on the composition of the corpus.

date back to the 1930s. One instance, reported in (18) below but not included in Figure 2, was found in 1928 with the verb *lick* used in the sense of 'beat, defeat'.

(18) *Swap generals with us and we'll lick the hell out of you.*

This suggests that the construction was present (although perhaps less common) before the 1930s. This is confirmed by a quick survey in the American portion of the much larger Google Books n-gram corpus (Davies 2011), where the *hell*-construction is first attested (though scarcely) in the 1910s and 1920s, and undergoes a sudden rise in frequency in the 1930s.

At any rate, plot (A) in Figure 2 shows that the construction has been steadily increasing in frequency since its arrival in the language. Also, more and more different verbs are attested in the construction, as seen by the increase in type frequency in plot (B). Because the type frequency measured in a corpus depends to a large extent on the token frequency of the relevant construction, it is also pertinent to relativize the increase in type frequency by calculating the token/type ratio, which is also a common measure of morphological productivity (Baayen and Lieber 1991). Except for two sudden declines in the 1950s and in the 1990s, the token/type ratios also point to a general increase in the scope of the construction, as seen in plot (C).

The increase in type frequency and token/type ratio reflects an expansion of the productivity of the construction, but it does not show the structure of this productivity. For instance, it does not say what kinds of verbs joined the distribution (and when), whether there are particular semantic domains preferred by the construction, and whether and how this changes over time. To answer these questions, this study analyzes the distribution of the construction from a semantic point of view by using a measure of semantic distance derived from distributional information. The distributional semantic model is described in Section 5.3, and evaluated in Section 5.4. In Section 5.5, two visualization techniques are presented that use this model to investigate the diachronic development of the semantic distribution of the construction. In Section 5.6, the diachronic data is submitted to statistical analysis to evaluate how the semantic structure of the distribution predicts how verbs are productively used in the construction.

5.3 The vector-space model

One of the goals of this case study is to assess the structure of the semantic domain of the *hell*-construction at different points in time, using measures of semantic distance derived from distributional information. To achieve this, we need to

obtain naturally occurring instances of all verbs attested in the construction from a large corpus, in their context of use. Various corpora of sufficient size for vector-space semantic modeling are available, some of which are commonly used for that purpose: for instance, the 100 million-word British National Corpus, the two billion-word ukWaC corpus of blogs from the .uk domain, and Wikipedia dumps. This study uses the COCA, because as a well-balanced corpus of American English it is more ecologically valid for this study than the other cited resources, which consist of a different variety of English and/or are more genre-specific. 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, an even more ecologically valid choice would have been to use data from a particular time frame to build a vector-space model for the analysis of the distribution of the construction in the same time frame. However, it did not prove possible to find enough data to achieve that purpose, since even the twenty or so million words per decade from the COHA turned out to be insufficient to assess the meaning of words from their distribution with a reasonable degree of reliability. Using data from the 1990s and 2000s to model the semantics of lexemes used in earlier decades is actually not as problematic as it might sound, since the meaning of the verbs under consideration are not likely to have changed considerably within the relatively short and recent time frame of this study, in which American English had long been standardized and its semantics (just like its grammar) regulated by such authoritative sources as Webster's *American Dictionary of the English Language*, whose first edition had been published in 1828.[8] Besides, using the same distributional data entails that a common semantic space will be used for all time periods, which makes it easier to visualize changes.

All instances of the relevant verbs were extracted from the online version of the COCA with their context of occurrence. Verbs judged not frequent enough to assess their meaning from their distribution (i.e., less than 2,000 occurrences) were excluded from the study: *bawl*, *belt*, *cream*, *dent*, *disgust*, *flog*, *grease*, *horsewhip*, *infilade*, *infuriate*, *irk*, *lam*, *micromanage*, *mortgage*, *nag*, *nuke*, *sodomize*, and *squash*. This left a total of 92 usable verbs. The words in the sentence contexts extracted from the COCA were lemmatized and annotated for part-of-

8 This is not to say that semantic changes cannot occur within this time frame; after all, there have been major social, cultural, and technological changes since the 1930s that are most likely to be reflected in language. Both Boussidan (2013) and Gulordava and Baroni (2011) detect semantic change in distributional data for particular words within much shorter time spans. Semantic change should however be minimal since the 1930s for the verbs considered in this study, especially as far as the similarity between them is concerned.

speech using TreeTagger (Schmid 1994).[9] The matrix of co-occurrences between the target verbs and their lemmatized collocates within a five-word window was computed on the basis of the annotated data, as described in Section 4. Tokens with the same lemma and a different part of speech (e.g., the noun *place* as in *dinner at my place*, and the verb *place* as in *place the envelope in the printer tray*) were considered different collocates and, accordingly, received a different frequency count. Only the noun, verb, adjective, and adverb collocates listed among the 5,000 most frequent words in the corpus were considered (to the exclusion of the verbs *be*, *have*, and *do*),[10] 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).[11] The resulting matrix, which contains the distributional information for 92 verbs occurring in the *hell*-construction, constitutes the semantic space under consideration in this case study. The rest of the analysis was conducted on the basis of this semantic space in the R environment (R Development Core Team 2013).

5.4 Evaluation of the vector-space model

Before turning to the analysis of the *hell*-construction proper, the validity of the vector-space model to capture semantic similarity between verbs is first evaluated. To visualize similarity relations and possible groupings that can be inferred from the distributional data, the rows of the co-occurrence matrix were submitted to hierarchical clustering, using the "hclust" function of the R environment. Hierarchical clustering is an unsupervised learning technique aimed at the classification of a set of objects into homogenous categories (cf. Aldenderfer and Blashfield 1984), according to a set of numerical variables against which each object (here, each verb) is characterized. In the present case, the variables are the weighted co-occurrence counts recorded in each row of the matrix, and two rows are considered more similar if they have similar co-occurrence counts in the same columns, which was measured by the cosine distance, using the "cosine" function from the R package lsa (Wild 2007). The hierarchical clustering algorithm uses pairwise distances between rows to recursively merge the two most similar observations or clusters of observations into a higher-level cluster, until there is only one cluster containing all objects. The distance between clusters depends on

9 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (consulted Feb 7 2014).

10 The list of the 5,000 most frequent words in the COCA was downloaded from http://www.wordfrequency.info/free.asp (consulted Feb 7 2014).

11 http://clic.cimec.unitn.it/composes/toolkit/ (consulted Feb 7 2014).

which linkage criterion is used. For this study, Ward's linkage method was chosen. Ward's criterion aims to minimize the variance within clusters when choosing which two clusters to merge, and, compared to other linkage methods, it has the property of generating more "compact" clusters.

The output of the hierarchical clustering algorithm is thus, as the name indicates, a hierarchy of clusters. This hierarchy is generally presented in the form of a tree diagram, or *dendrogram*, in which the observations are leaves, and the clusters are branches linking the observations at different levels. The dendrogram resulting from the cluster analysis of the 92 verbs found in the *hell*-construction between 1930 and 2009 is presented in Figure 3. This kind of

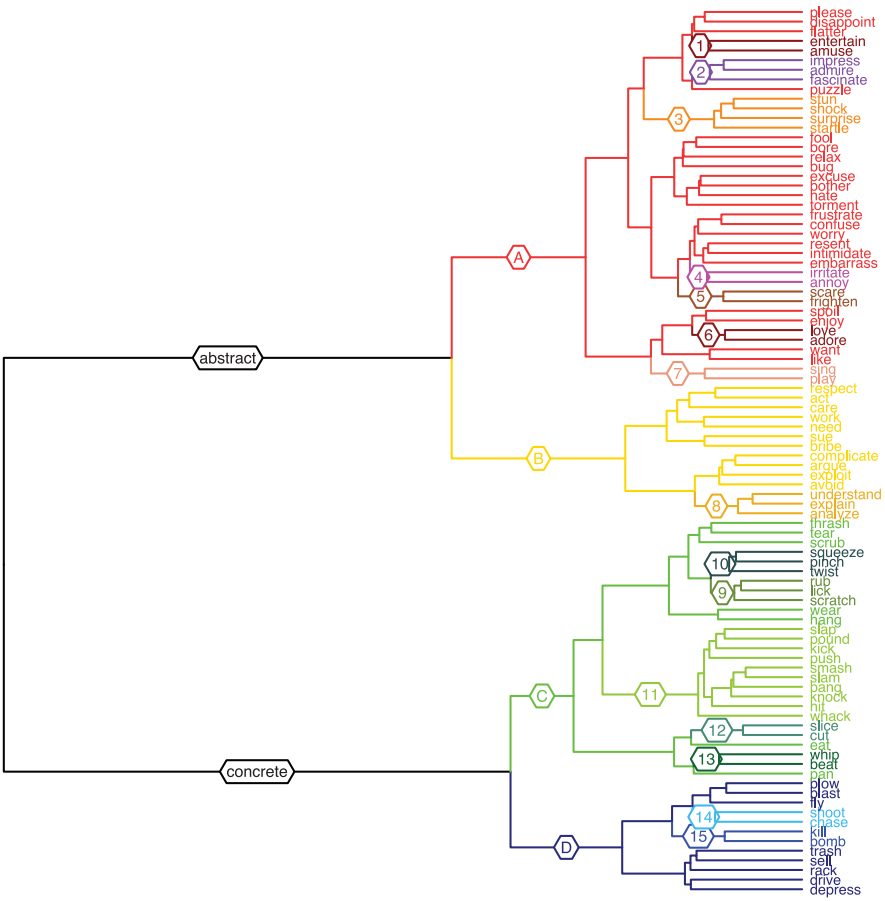

Figure 3: Cluster dendrogram for all verbs in the distribution of the *hell*-construction between 1930 and 2009.

diagram arranges items according to their similarity, and as it were, traces the history of cluster mergers by the algorithm, from the earliest ones on the rightmost side of the graph, to the last one on the leftmost side. Clusters located towards the top of the tree (here on the right) represent tight groupings of highly similar items, while clusters located at lower levels (here on the left) correspond to looser groups of items related by more abstract commonalities.

A number of highly meaningful groupings can be found in the dendrogram; they are indicated by branch labels and different color codings in Figure 3. First, there are several pairs of synonyms or near-synonyms that were merged together as a single cluster before being linked to other clusters. In other words, these words were nearest neighbors in the semantic space: from top to bottom, *entertain* and *amuse* (1), *irritate* and *annoy* (4), *scare* and *frighten* (5), *love* and *adore* (6), *slice* and *cut* (12), and *whip* and *beat* (13). *Love* and *adore* are joined by *like* and *enjoy* in a higher-level cluster. There are also several groups containing words that clearly relate to the same notional domain without necessarily being synonyms: *impress*, *admire*, and *fascinate* from group (2) lexicalize feelings of awe, *stun*, *shock*, *surprise*, and *startle* from (3) relate to astonishment, and *sing* and *play* (7) are verbs of performance. At a yet more abstract level, we find that some verbs seem to fall in the same cluster because they share some abstract property related to their experiential Gestalt and/or other entities typically involved in the event they describe. In group (8), *understand*, *explain*, and *analyze* relate to mental processes and typically involve complex and abstract ideas. In (9), *squeeze*, *pinch*, and *twist* share the notion of contact and exertion of a force, typically with the hand or fingers, and in (10), *rub*, *lick*, and *scratch* involve repeated motion on a surface. The two clusters are merged at a higher level, presumably because they are unified by the notion of contact. *Shoot* and *chase* in (14) correspond to different aspects of hostile engagement with a target (like in hunting), and *bomb* and *kill* in (15) are both violent and harmful to human beings. Finally, (11) is a large cluster that contains almost all verbs of hitting (especially in a violent way) found in the distribution: *slap*, *pound*, *kick*, *smash*, *slam*, *bang*, *knock*, *hit*, *whack*, plus *push*, which also contains a similar force component. This group constitutes a coherent semantic class that can evidently be derived from distributional information.

At a higher level, the clustering algorithm partitions the distribution in a way that also makes intuitive sense. As indicated in Figure 3, the verbs are neatly divided into concrete and abstract domains, i.e., verbs primarily describing physical actions vs. verbs describing actions that do not have a clear concrete manifestation, such as feelings, emotions, mental processes and other

abstract events. The two types of verbs are further divided into four semantically coherent groups, labeled A to D in Figure 3. Group (A) mostly contains psych-verbs describing feelings and emotions: *please*, *surprise*, *hate*, *worry*, *annoy*, *like*, etc. Group (B) contains the other kinds of abstract actions. Group (C) mostly contains physical actions that typically involve contact and exertion of a force on a second participant, resulting in an effect that is often damaging: *scrub*, *slap*, *push*, *whack*, *cut*, *beat*. The verbs in group (D) have a weaker force component (if at all), and do not necessarily involve contact. A few of them describe perfectly harmless actions that do not have the causative character of the verbs in group (C), e.g., *drive*, *sell*.

More could be said about the cluster analysis reported in Figure 3 and the semantic distributional model it is based on, but the comments made so far already amply illustrate that the measure of semantic similarity provided by this vector-space model accurately reflects semantic intuitions. This is not to say that the model never makes mistakes or would not enter in disagreement with human speakers as to what verbs are more similar to each other, as indeed there seems to be a few misclassifications. For example, *want*, but not *enjoy* or *love*, turns out as the nearest neighbor of *like* (contrary to intuition), and *depress* is grouped with verbs of physical actions. Such mistakes occur when a word shares more of its distribution with words that are not truly similar to it than with words that are, and could possibly be avoided by relying on a finer notion of word context (for instance by taking into account grammatical dependencies, cf. Padó and Lapata 2007). Be that as it may, this distribution-based measure of semantic similarity is on the whole highly satisfactory, which warrants its use for the study of the syntactic productivity of the *hell*-construction.

5.5 Visualizing productivity with 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. Taking up the conception of meaning as a space populated by words which lexicalize portions of it, a distance measure can be seen as providing a way to locate words in that space with respect to each other. By aggregating the semantic distance information for all items in the distribution, we can form an impression of the structure of the semantic domain of the construction, which can be given a visual depiction. In particular, a visual representation allows 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 the regions that are more sparsely populated (with fewer and/or more scattered verbs). In turn, by observing the structure of the semantic domain of the construction at different points in time, we can gain insights into the diachronic development of its productivity. This section presents an analysis of the *hell*-construction in diachrony that draws on two well-known visualization techniques to identify patterns in the semantic distribution.

The first of these techniques, multidimensional scaling (MDS), provides a way both to aggregate distance information and to represent it visually. It aims at placing objects in a space with (usually) two dimensions such that the between-object distances are preserved as much as possible. Each object is assigned coordinates by the algorithm, which can be used to generate a plot that visually depicts the similarity relations between objects. The pairwise distances between all verbs in the distribution were submitted to multidimensional scaling into two dimensions, using the "isoMDS" function from the MASS package in R. This is essentially tantamount to mapping the high-dimensional distributional space of the co-occurrence matrix, where each of the 4,683 collocates is one dimension, into a two-dimensional space, which should offer an approximation of the semantic space. The new distance matrix computed in the resulting two-dimensional space displays a good level of correlation with the original distance matrix (Pearson's $r = 0.8295$, $t(4184) = 96.0795$, $p << 0.001$), and the algorithm returns a satisfactory stress value (Kruskal's stress = 0.2017, cf. Kruskal 1964). This shows that the two-dimensional space returned by MDS is reasonably faithful to the original high-dimensional space.

To visualize changes in the semantic domain of the *hell*-construction, the diachronic data set was divided into four successive twenty-year periods: 1930–1949, 1950–1969, 1970–1989, and 1990–2009. This division was chosen for purely practical reasons: the corpus is sampled by decades, but decades turned out to be too short timespans to observe significant changes from one period to another. The distribution of the construction in each time period was extracted and plotted in a separate graph, using the set of coordinates returned by MDS. These *semantic plots* are presented in Figure 4. For convenience and ease of visualization, the verbs are color-coded according to the four broad semantic groupings that were identified by the cluster analysis presented in Section 5.4 (cf. Figure 3). For the sake of comprehensiveness, token frequency is also represented in the plots, although it will not be subsequently discussed. Verbs with a token frequency greater than one are plotted with a circle in addition to their label; the size of the circle is proportional to the natural logarithm of the token frequency.

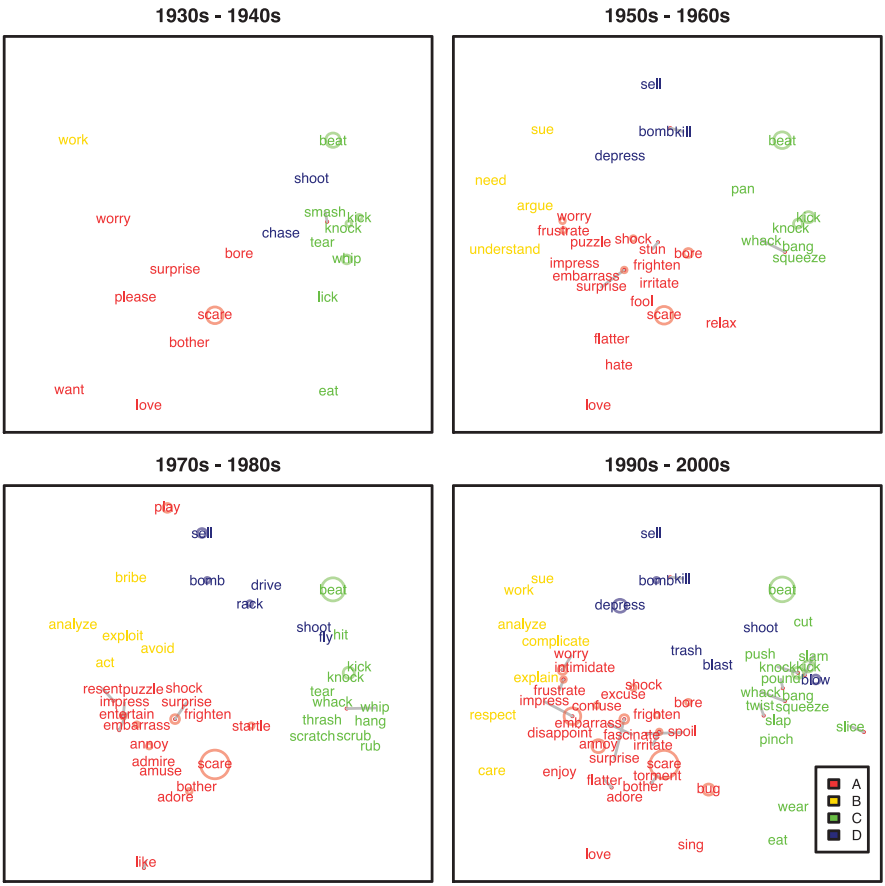

Figure 4: Semantic plots of the *hell*-construction in four successive twenty-year periods. The colors correspond to the four clusters of verbs identified by cluster analysis (cf. Figure 3).

By comparing the plots in Figure 4, we can follow the semantic development of the *hell*-construction.[12] First, one thing that is particularly striking is that the construction is clearly centered around two kinds of verbs: psych-verbs

12 Note that this idea and its technical implementation are similar to the concept of motion charts, proposed by Hilpert (2011) to visualize linguistic change on the basis of the frequency of co-occurrence of lexical and grammatical features. The semantic plots showcased in this paper differ from Hilpert's motion charts in two respects: (i) they are exclusively based on co-occurrences with lexical items as a means to represent semantic relations, and (ii) they are designed to visualize not how grammatical patterns of usage change over time (as reflected by the motion of items in the plotted space), but how the semantic domain of a construction is filled at different points of its history.

(*surprise*, *please*, *scare*, etc.) and verbs of hitting (*smack*, *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 the lion's share of the distribution at the outset, and they continue to 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.

These observations illustrate the role of type frequency and variability in productivity. In the semantic space, densely populated regions appear to be the most likely to attract new members. However, this observation is derived by informally eyeballing the data, and is not the result of a systematic and principled analysis. Besides, one problem with MDS is that it 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. Even though the results of MDS received good measures of reliability, some distortion can indeed be seen in Figure 4 in the spatial overlap between the groupings that were identified by cluster analysis (for instance *blow*, a member of group D, is placed among verbs of group C), as well as in the fact that some verbs are clearly misplaced. For instance, *play* and *sing* are positioned far apart from each other, while they were identified as nearest neighbors in the high-dimensional space. In sum, even though MDS is decidedly useful for the purpose of exploratory analysis, the semantic plots it generates should be taken with a grain of salt and its results be compared with another, more reliable method.

To analyze how regions of the semantic space fill up over time, verbs can be grouped according to cluster analysis, as diagrammed in Figure 3. If the dendrogram of Figure 3 is cut at a given level, a list of clusters of comparable internal coherence can be obtained. By combining a given clustering solution with the diachronic partition of the distribution into periods created to construct the semantic plots, the variation in the size of each cluster over time can be plotted. This is presented in Figure 5. Each plot charts the number of verbs in each cluster for four clustering solutions, respectively containing three, four, five, and six clusters. The exact nature of these clusters and the verbs they contain can be determined from the dendrogram in Figure 3, but this information is not necessary in order to observe that it is always the groups containing

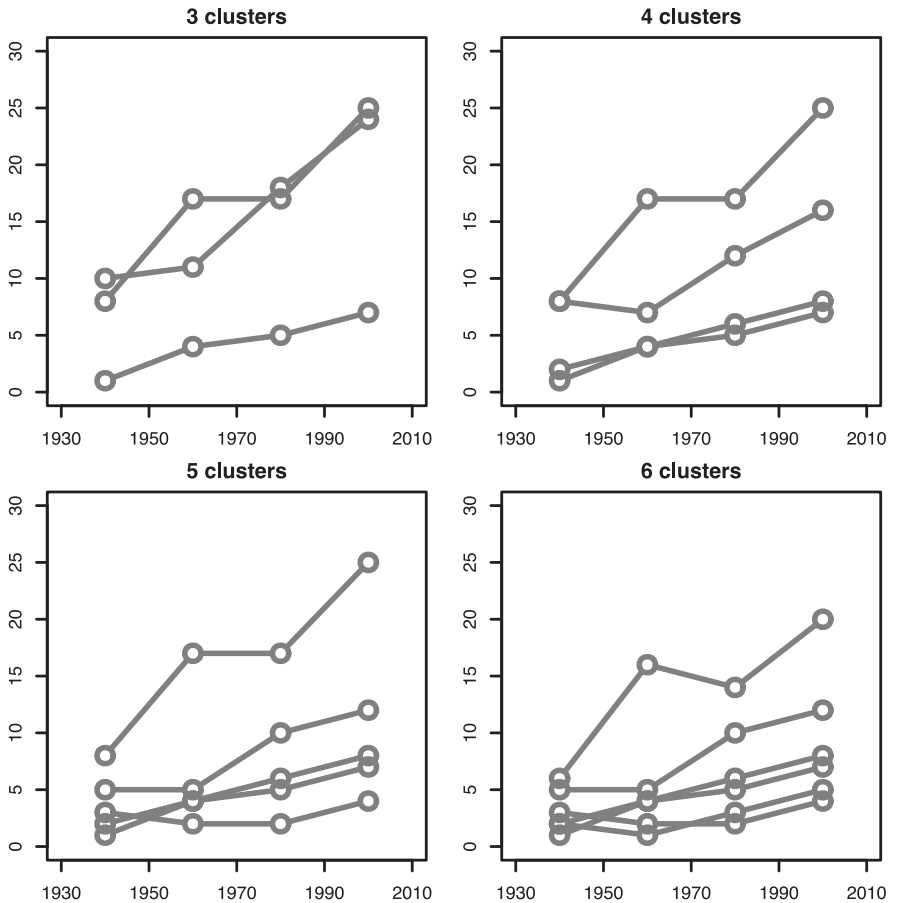

Figure 5: Type frequency variation of semantic clusters at different levels of granularity (viz., different numbers of clusters).

the most members at the outset that most quickly gain new members afterwards. There is indeed a high and significant correlation between the initial size of a cluster and its mean growth, i.e., the mean increase in size of the cluster in later periods, across the four clustering solutions (Pearson's $r = 0.8024$, $t(16) = 5.3786$, $p < 0.0001$). Two clusters in particular stand out as the front runners in the productivity of the construction at any degree of granularity; unsurprisingly, they correspond more or less to the two semantic classes described above (psych-verbs and forceful actions). In sum, the results of cluster analysis are largely in line with the semantic plots, confirming previous observations with a more principled and reliable method. These findings illustrate one aspect of the

role of type frequency in productivity: within a given semantic class, new items are more likely to appear if many items are already present. Outside of the two identified domains of predilection, other classes never become important because they do not receive a "critical mass" of items, and therefore attract new members more slowly.

At the same time, there is a notable difference between forceful action verbs and psych-verbs that is more visible in the semantic plots than in the dendrogram: the former always form a tight and compact cluster, while the latter occupy more space. This is coherent with the intuition that the two categories differ in the variety of situations that they can cover: the types of hitting and other forceful actions are rather limited, but the range of feelings and emotions experienced by humans is more varied. The two clusters start with a similar number of verbs, but they have a different structure, in that psych-verbs are more scattered, leaving gaps that are gradually filled over time. I would argue that it is a crucial difference that accounts for why psych-verbs turn out to be more productive than forceful action verbs, despite having the same starting type frequency, in line with the idea that semantic variability promotes productivity. Importantly, this finding illustrates another useful aspect of semantic plots: compared to cluster dendrograms, they allow to better appreciate how items are spread in the semantic space, and not only how they cluster together, and to visualize the shape of clusters, and not only their size.

5.6 Statistical analysis

The visualization techniques described in the previous section (multidimensional scaling and cluster analysis) prove useful to explore and analyze the productivity of constructions on the basis of distributional semantic data. There is, however, more to be offered by the distributional quantification of meaning for the study of syntactic productivity. In particular, one major advantage presented by a quantification of meaning over an introspective approach is that it allows measures capturing particular aspects of the semantic space to be derived and used in statistical analysis. In this section, it is shown that a measure of density derived from a distributional semantic space is a significant predictor of syntactic productivity.

The quantitative analysis presented in this section is based on the following premises. Given that the *hell*-construction conveys essentially the same meaning since its inception, all verbs ever attested in the construction form equally plausible combinations with it from a semantic point of view. However, they are clearly not all equally likely to occur at any point in the history of the

construction, as shown by the diachronic data presented in the last section. According to contemporary usage-based accounts of syntactic productivity, the probability of a new coinage depends on properties of the prior usage of the construction (cf. Section 2), especially as it relates to the presence of similar items in the attested distribution. In diachrony, this usage-based account translates into the expectation that the usage of a construction at a given point in time should determine its distribution at a later point (at least partly). More precisely, a given item is not likely to join the distribution of the construction until a particular set of conditions are met. This prediction can be tested by determining if there is a relation between the probability that a given verb will join the distribution in a given period of the history of the *hell*-construction, and the structure of the semantic domain of the construction in the immediately preceding period. In particular, it is suggested in this section that the occurrence of a new item in the construction is related to the density of the semantic space around this item. This notion of density can be seen as an operationalization of the concept of coverage suggested by Suttle and Goldberg (2011) to explain their experimental results (cf. Section 2).[13]

For each verb in the distribution, the period of first occurrence in the construction was determined. For the verbs first occurring in 1970–1989 and 1990–2009, the binary variable Occurrence was set as true for the first period of occurrence, and as false for all earlier periods (later periods were ignored). The verbs first occurring before 1970 could not be included in the analysis for two logical reasons. For verbs first occurring in 1930–1949, there is no earlier period from which to extract a measure of density. For verbs first occurring in 1950–1969, there is no period of non-occurrence with which to establish a comparison, because that period would be 1930–1949, which, as just pointed out, cannot receive a measure of density.

For each verb-period pair thus obtained, a measure of density was computed that captures how populated the semantic space was in the neighborhood of the verb in the immediately preceding period. For instance, *explain* is first attested in the construction in the fourth period (1990–2009); the variable Occurrence is thus true for Verb = *explain* and Period = 1990–2009, and the measure of density is computed on the semantic space from the third period (1970–1989). Two other

13 As pointed out by one of the anonymous reviewers, an operationalization of Barðdal's (2008) concept of semantic coherence could also be similarly arrived at using distributional measures of semantic similarity. However, contrary to the notion of density described here, a measure of semantic coherence would not be defined relative to a particular point of the semantic space where a potential new coinage is considered, but would be a property of a construction as a whole.

data points are added for *explain* with PERIOD = 1950–1969 and PERIOD = 1970–1989, with OCCURRENCE set to false, and the density measures are respectively calculated from the semantic spaces of 1930–1949 and 1950–1969. I used mixed effects logistic regression to determine if there exists a quantitative relation between the measure of density and the probability of first occurrence of the verb in the construction.

One of the main questions to be addressed is how to measure the density of the semantic space at a given point in this space (corresponding to a particular verb). The measure of density should take both the number of neighbors and their proximity into account, in that it should capture to what extent a large number of items are found in the close vicinity of that point. Also, the measure of density should be defined locally, i.e., it should consider a limited portion of the semantic space and by no means all of it, otherwise it will not be a good predictor of the probability of a new coinage, since it will invariably increase with the sheer number of attested items in the entire space, regardless of how relevant these items are (see Suttle and Goldberg 2011: 1243 for a similar observation). In this study, I suggest a measure of density that considers the set of the N nearest neighbors of a given item in the semantic space. This measure of density is defined by the following formula:

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

where $d(X, Y)$ is the distance between two items X and Y, and V_n is the n-th nearest neighbor of V. In plain language, the density around a given item V is equal to one minus the mean distance of the N neighbors to this item. The mean distance to nearest neighbors decreases with the density of the surrounding space, 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. For instance, for $N = 3$, the density of the semantic space around a verb V that has V1, V2 and V3 as nearest neighbors at the respective distances of 0.3, 0.4, and 0.5 amounts to $1-(0.3 + 0.4 + 0.5)/3 = 0.6$.

The dataset was used to fit a linear mixed effects model using the function "lmer" from the lme4 package (version 1.0-5) in the R environment (Bates et al. 2011). In this model, OCCURRENCE is the dependent variable, and the measure of density is the single predictor. As for random effects, the model also includes by-verbs random intercepts and random slopes for DENSITY. This was done in order to factor in variation in density related to individual verbs; recall that what we want to test is whether the first occurrence of a new verb is heralded by an increase in the density of the semantic space around that verb. Different versions of the density measure were calculated by considering different numbers of

nearest neighbors between 3 and 8 (the N variable in the formula). The predictive power of each version of the density measure was tested in a different model. The results of these models are summarized in Table 2.

Table 2: Results of mixed effects logistic regression models predicting the first occurrence of a verb in the *hell*-construction from measures of semantic density based on 3 to 8 nearest neighbors. Model formula: Occurrence ~ Density + (1 + Density | Verb)

Nearest neighbors (N)	Effect of Density	*p*-value	significant?
3	0.7211	0.195	no
4	0.8836	0.135	no
5	1.0487	0.091	marginally
6	1.2367	0.056	marginally
7	1.4219	0.034	yes
8	1.6625	0.017	yes

For all values of N, a positive effect of Density is found. In other words, a higher space density positively increases the odds that a new verb occurs in the construction. However, the effect is only significant for $N \geq 7$; more generally, the *p*-value decreases as N increases. In sum, the effect of density is both stronger and more robust when a larger number of neighbors is considered in its calculation. The variation in effect strength indicates that 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. The variation in *p*-value means that the relation between Density and Occurrence is not as systematic when Density is measured on fewer neighbors. I would argue that this fact is another manifestation of the role of type frequency in syntactic productivity: a measure of density that is supported by a higher number of types makes more reliable predictions than a measure supported by fewer types. This means that productive uses not only depend on whether the meaning of a potential coinage is sufficiently related to the existing semantic space, they also occur more reliably when this relation is supported by many items. These findings point to complementary roles that the semantic distribution of a construction and its type frequency play in syntactic productivity, as argued by Barðdal (2008): the former sets the necessary conditions for a new coinage to occur, while type frequency increases the confidence that this coinage is indeed possible.

As one reviewer points out, it should be acknowledged that the effect of density appears to be partly in contradiction with my earlier observation from Figure 4 that the tight cluster of forceful action verbs at the outset (1930s–1940s)

turns out to be less productive in later periods than the looser cluster of psych-verbs, since the former indicates higher semantic density than the latter, and yet does not seem to attract as many new items. For reasons previously explained, the statistical models summarized in Table 2 were trained on data for the verbs occurring in the third and fourth period (1970–1989, 1990–2009) for the first time; hence, they are not affected by this discrepancy. That being said, this seems to indicate that the density-based model only provides a partial account of the productivity of the construction. The situation found for the *hell*-construction, which was given an intuitive explanation, is in line with Suttle and Goldberg's (2011: 1253) finding that a coinage located next to a tight cluster (low variability, moderate similarity) is judged less acceptable than one located next to a loose cluster (high variability, moderate similarity). Both findings suggest that productivity is not sensitive to the mere presence of similar items, but also, and perhaps more importantly, to the distribution of these items in the semantic space, which the present measure of density does not capture.

It should be insisted that the method presented in this section is only intended to illustrate how statistical analysis can be applied to the study of syntactic productivity in the approach advocated in this paper. As such, the coarse density-based models described above are obviously not meant to capture the phenomenon in its full complexity, but merely to identify a general trend. In future research, other, more complex measures of density capturing different aspects of the semantic space should be devised, and other factors tested in order to identify the determinants of syntactic productivity on a quantitative basis. In the meantime, it is hoped that the present work demonstrates the methodological potential of a distributional approach to semantics for such studies.

6 Conclusion

This paper presents the first attempt at using a distributional measure of semantic similarity for the study of syntactic productivity in diachrony, i.e., the property of the slots of grammatical constructions to attract new members over time, thus extending their distribution. According to contemporary accounts of syntactic productivity, speakers tend to use constructions with items that are semantically similar to the previously attested items, and only depart from the established semantic domain if there is already some variability in the distribution (Barðdal 2008; Bybee and Eddington 2006; Suttle and Goldberg 2011). Crucially, these accounts rely to a large extent on semantics, especially with respect to how a potential new item semantically relates to the existing

distribution. Consequently, testing these theories on empirical data necessitates an operationalization of the meaning of words in general, and of semantic similarity in particular, which raises methodological issues.

Neither of the two existing approaches to the operationalization of meaning is entirely satisfactory: using the semantic intuitions of the linguist raises issues of objectivity and mainly produces categorical data, from which it is not possible to directly derive measures of similarity and variability, while collecting judgements of similarity from native speakers raises problems of scalability, in that it is practically feasible only when a limited number of items are considered. In this paper, a third alternative was considered that avoids the limitations of both kinds of approaches, consisting in using distributional information as a proxy to word meaning. Drawing from the observation that words with a similar meaning tend to have similar collocates, it is possible to base a measure of semantic similarity on co-occurrence information derived from large corpora. The measure of semantic similarity provided by so-called vector-space models of word meaning usually compares well to human semantic intuitions, and presents the advantage of being entirely data-driven.

In a case study of the construction "V *the hell out of* NP" (e.g., *You scared the hell out of me*) in American English, it was shown how the distributional semantic approach to semantic similarity can be applied to the study of syntactic productivity in diachrony. Multidimensional scaling and cluster analysis were used as means of visually representing the distributional semantic information provided by a vector-space model. It was shown how these visualization techniques can be used to identify the semantic domains preferred by the construction, and to plot its semantic evolution in four successive 20-year periods from 1930 to 2009. The results of this exploratory analysis were in line with current views on the determinants of syntactic productivity. Finally, the data were submitted to statistical analysis using mixed effects logistic regression, which revealed a positive effect of the density of the semantic domain of the construction around a particular item on the probability that this item will join the construction in the next time period. This finding is compatible with Suttle and Goldberg's (2011) notion of coverage as a driving factor of productivity, which indicates the relevance of this notion to diachronic data. Moreover, it was also found that the robustness of this effect increases with the number of items that are considered in the calculation of the density measure. This finding points to a complementary role of type frequency, which increases the confidence that a particular coinage is possible.

In sum, the present study demonstrates that a distributional approach to meaning not only provides an appropriate measure of semantic similarity, it also enables the use of methods for which quantification is necessary, such as data

visualization and statistical analysis. That being said, this paper has only scratched the surface of what this method can accomplish, and the range of other questions it could address is yet to be explored. In particular, it could allow for testing the influence of different aspects of the semantic space (beyond density) on productivity, which the case of the *hell*-construction did not exemplify, such as the interaction between semantic similarity and token frequency (Bybee 2010). In conclusion, distributional semantics is a promising approach for the study of syntactic productivity, and possibly for other domains where semantic similarity is relevant.

References

Aarts, Bas, Joanne Close, Geoffrey Leech & Sean Wallis (eds.). 2013. *The verb phrase in English: Investigating recent language change with corpora*. Cambridge: Cambridge University Press.

Andrews, Mark, Gabriella Vigliocco & David P. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116(3). 463–498.

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. Amsterdam & Philadelphia: John Benjamins.

Barðdal, Jóhanna. 2011. Lexical vs. structural case: A false dichotomy. *Morphology* 21. 619–654.

Boussidan, Armelle. 2013. *Dynamics of semantic change: Detecting, analyzing and modeling semantic change in corpus in short diachrony*. Lyon: Université Lumière Lyon 2 dissertation.

Bowerman, Melissa. 1988. The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In John A. Hawkins (ed.), *Explaining language universals*, 73–101. Oxford: Blackwell.

Baayen, Harald & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29. 801–844.

Bates, Douglas, Martin Maechler & Ben Bolker. 2011. *lme4: Linear mixed-effects models using S4 classes*. R package. http://CRAN.R-project.org/package=lme4 (accessed 21 February 2014).

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam & Philadelphia: John Benjamins.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*,49–69. Oxford: Oxford University Press.

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

Bybee, Joan & Sandra Thompson. 1997. Three frequency effects in syntax. *Berkeley Linguistics Society* 23. 65–85.

Chomsky, Noam. 1986. *Knowledge of language*. Cambridge, MA: MIT Press.

Clausner, Timothy C. & William Croft. 1997. Productivity and schematicity in metaphors. *Cognitive Science* 21(3). 247–282.

Cook, Paul & Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta*, 28–34.

Croft, William. 2010. Relativity, linguistic variation and language universals. *CogniTextes* 4. http://cognitextes.revues.org/303 (accessed 7 February 2014).

Davies, Mark. 2008. *The corpus of contemporary American English: 450 million words, 1990–present*. http://corpus.byu.edu/coca/ (accessed 7 February 2014).

Davies, Mark. 2010. *The corpus of historical American English: 400 million words, 1810–2009*. http://corpus.byu.edu/coha/ (accessed 7 February 2014).

Davies, Mark. 2011. *Google books corpus. (based on Google books n-grams)*. http://google books.byu.edu/ (accessed 7 February 2014).

Dewey, Tonya Kim. 2006. *The origins and development of Germanic V2: Evidence from alliterative verse*. Berkeley: University of California Berkeley dissertation.

Dinu, Georgiana, Nghia The Pham & 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*), 31–36. East Stroudsburg, PA: ACL.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60.

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

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (Special volume of the Philological Society), 1–32. Oxford: Blackwell.

Glenberg, Arthur M. & David A. Robertson. 2000. Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* 43(3). 379–401.

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

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, Adele E. 2011. Meaning arises from words, context, and phrasal constructions. *Zeitschrift für Anglistik und Amerikanistik* 59(4). 331–346.

Goldberg, Adele E. & Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80(3). 532–568.

Grefenstette, Gregory. 1994. *Explorations in automatic thesaurus discovery*. Dordrecht: Kluwer.

Gries, Stefan Th. & Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*,73–90. Stanford, CA: CSLI.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*, 67–71. East Stroudsburg, PA: ACL.

Haïk, Isabelle. 2012. *The hell* in English grammar. In Nicole Le Querler, Franck Neveu & Emmanuelle Roussel (eds.), *Relations, connexions, dépendances: Hommage au professeur Claude Guimier*,101–126. Rennes: Presses Universitaires de Rennes.

Hilpert, Martin. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics* 16(4). 435–461.

Hock, Hans H. & Brian D. Joseph. 1996. *History, language change and language relationship. An Introduction to historical and comparative linguistics*. Berlin & New York: Mouton de Gruyter.

Hoeksema, Jack & Donna J. Napoli. 2008. Just for the hell of it: A comparison of two taboo-term constructions. *Journal of Linguistics* 44(2). 347–378.

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

Krug, Manfred. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin & New York: Mouton de Gruyter.

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

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25(2–3). 259–284.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

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

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University Of Chicago Press.

Levshina, Natalia & Kris Heylen. 2014. A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.), *Extending the scope of Construction Grammar*, 17–46. Berlin & Boston: De Gruyter Mouton.

Lund, Kevin, Curt Burgess & Ruth A. Atchley. 1995. Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings (LEA)*, 660–665.

Mair, Christian. 2002. Three changing patterns of verb complementation in Late Modern English: A real-time study based on matching text corpora. *English Language and Linguistics* 6(1). 105–131.

Mair, Chistian. 2006. *Twentieth-century English: History, variation and standardization*. Cambridge: Cambridge University Press.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

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

Padó, Sebastian & Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2). 161–199.

Pedersen, Ted. 2006. Unsupervised corpus-based methods for WSD. In Eneko Agirre & Philip Edmonds (eds.), *Word sense disambiguation: Algorithms and applications*,133–166. Dordrecht: Springer.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press/Bradford Books.

Pantel, Patrick & Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada*, 613–619.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–124.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Purandare, Amruta & Ted Pedersen. 2004. Word Sense Discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL), May 6–7, 2004, Boston, MA*, 41–48.

R Development Core. 2013. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/ (accessed 21 February 2014).

Ross, John R. 1973. Nouniness. In Fujimura Osamu (ed.), *Three dimensions of linguistic research*. Tokyo: TEC Company Ltd.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics, Athens, Greece*, 104–111.

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

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester*, 44–49.

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

Traugott, Elizabeth C. & Graeme Trousdale. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press.

Turney, Peter & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.

Vyas, Vishnu & Patrick Pantel. 2009. Semi-automatic entity set refinement. In *Proceedings of NAACL-09, Boulder, CO*, 290–298.

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.

Wild, Fridolin. 2007. An LSA package for R. In Fridolin Wild, Marco Kalz, Jan van Bruggen & Rob Koper (eds.), *Mini-Proceedings of the 1st European Workshop on Latent Semantic Analysis in Technology-Enhanced Learning, Heerlen, NL*, 11–12.

Wonnacott, Elizabeth, Jeremy K. Boyd, Jennifer Thompson & Adele E. Goldberg. 2012. Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language* 66(3). 458–478.

Zeschel, Arne. 2012. *Incipient productivity: A construction-based approach to linguistic creativity*. Berlin & Boston: De Gruyter Mouton.