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

Perek, Florent

DOI:
10.1515/ling-2015-0043

Document Version
Early version, also known as pre-print

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

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

Florent Perek

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 that is based on similarity in distribution 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 pinpoint the determinants of syntactic productivity in naturally occurring data.

1. Introduction

Grammars are in constant change.[1] 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 (probably for prosodic reasons), 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.

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.

1 I am grateful to Catherine Diederich, Adele Goldberg, Eva Maria Vecchi, Sascha Wolfer, and Arne Zeschel for their comments on earlier versions of this work. This paper is based on material presented at the workshop "Muster und Bedeutung" in Mannheim, Germany, on July 16 2013.

(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 abstract 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 16th 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 19th 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 I report 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, I describe an alternative solution to this methodological problem that makes use of distributional information as a proxy to meaning in order to derive a measure of semantic similarity. In section 5, I apply it 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. I show 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 word formation processes 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 particular word behaves as a noun, which was first coined (at least to my knowledge) 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.

Syntactic productivity manifests itself in various areas of linguistic behavior. In language acquisition, it guides 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, syntactic productivity is also involved 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 corresponds 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 on this last aspect that I will focus in this paper. It is commonly assumed that the same underlying process is at work in all of these domains. There is indeed evidence that it is, at any rate, driven by essentially the same factors, although there might well 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, 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 (if not more) than 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 that account, 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 at the extreme ends of the continuum to two kinds of productivity 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.

While Barðdal's model does make the right predictions for the Icelandic case and argument structure constructions that she discusses, one factor that she does not explicitly discuss is the similarity of novel items to previous usage. Under the reasonable assumption that speakers tend to use constructions in similar ways to their priorly witnessed usage, this factor is expected to play a major role in productivity. As a matter of fact, 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 relies on analogical extensions (on the basis of 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. However, it is not clear from Bybee and Eddington's or Barðdal's findings, whether productivity with highly general constructions (viz. with high semantic variability) also relies on analogy, or occurs independently of the existence of similar items. In other words, does there always have to be a similar item in the distribution of a construction for a novel item to be allowed in it, or can a construction 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 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 (a unique particle for each construction). After each set of sentences, participants were asked to rate how likely another instance of the same construction with a different verb (the target) was. With this design, Suttle and Goldberg could systematically manipulate type frequency (by varying the number of different verbs in each set), variability (by choosing the stimuli verbs from the same vs. different semantic classes, as determined by Levin's [1993] classification) and similarity (low vs. moderate vs. high similarity, by choosing the target verb from a particular semantic class, respectively one represented in the stimuli set vs. a similar class, i.e. concrete actions for both stimuli and target, vs. an unrelated class, i.e., verbs of cognition for the target), and test the effect of each factor (and their interactions) on the acceptability ratings of novel sentences provided by the participants.

Suttle and Goldberg find that type frequency and variability each have an independent effect, and that they are also involved in a positive interaction, in that 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 find a main effect of similarity, in that the closer a coinage is to an attested utterance, the more likely 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 was high, the effect of variability was negative, in that subjects were more confident about the coinage when the distribution of the construction was less variable, which Suttle and Goldberg suggest is because with low variability, participants see more verbs from the same class attested in the construction (since type

frequency was held constant), and therefore get more evidence suggesting that any verb from this class may be used in the construction. With moderate similarity, there was 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 classes. However, when similarity was low, there was no effect of variability, which means that variability is irrelevant when the target bears no resemblance to any attested item.

One of the merits of Suttle and Goldberg's contribution is to emphasize the importance of semantic similarity as a criterion for productive uses of a construction. To explain their results (especially the complex interaction between variability and similarity), they 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). Coverage relates 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. I describe these issues 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. In this section, I discuss how meaning can be tackled in empirical research on productivity, and describe the methodological and practical issues that are raised (see also Zeschel 2012 for a similar discussion). I suggest an alternative that relies on distributional semantics, and in the remainder of this paper I show how it can adequately be used to study the productivity of constructions.

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 gives access to signifiers (or forms), and accessing the meaning of these forms requires a human interpreter. Since looking at 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 thus inescapable for the study of syntactic productivity from a diachronic perspective (which is the perspective I adopt 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 of a referent, to more subtle judgements such as word sense distinctions (like distinguishing the uses of *pull* used as a transitive verb vs. a verb of motion) and even more complex tasks like identifying instances of an abstract construction. Importantly, manual annotation primarily produces categorical data, as the judgements it is mainly based on consist in deciding to which category a given item belongs. 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. More generally, manual semantic annotation poses the methodological problem that it is based on the semantic intuitions of a single individual (with the possible intervention of lexicographic sources if s/he is not a native speaker), 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.[2]

Bybee and Eddington (2006) suggest an alternative approach which consists in using the results of a semantic norming study, i.e., similarity judgements collected from a large group of native speakers. In such a study, participants are presented with pairs of items and are asked to rate how similar they find these items on a given scale.[3] By pooling the data from all participants, a more objective (or at least intersubjective) measure of semantic similarity between words is obtained that is 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. It is, however, decidedly less convenient and may even be more time-consuming than manual annotation, not to mention that it necessitates access to a population of native speakers ready to provide semantic judgements (usually for a compensation).[4] More importantly, it is also inherently limited in scope in that it is constrained by the number of judgements 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 judgements 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 require 190 judgements); by way of comparison, 50 and 100 items, which even a moderately productive construction easily attains, respectively require 1,225 and 4,950 judgements (ideally per participant). In sum, while a norming

2 Admittedly, the issue can be addressed by assigning the task to several annotators and checking for agreement, albeit at the cost of an even greater dependence on time and human resources.

3 Zeschel (2012) uses a refined variant of this method that takes into account not only semantic similarity as such but also various kinds of semantic relations, such as antonymy, hyponymy, and metonymical shift.

4 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, which is growing increasingly popular among psychologists.

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, I would like to evaluate another possible solution to the problem of assessing semantic similarity that was already mentioned by some authors 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 (such as part of speech, lemma, or syntactic structure) to electronic corpora. 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, I show in this paper how distributional semantics and its main computational linguistic implementation, the vector-space model, can also be fruitfully applied to augment corpus data with information about lexical meaning in an automatic, data-driven way that dispense (at least to a large extent) with the need for human semantic intuitions. I describe this technique in the next section.

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.[5] Conversely, differences in the distributional profile of two words is 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 (though not necessarily amounting to) co-occurrence counts. The first step in creating a vector-space model is to build a co-occurrence matrix, with the set of words under consideration as rows, and the collocates against which

5 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 the same contexts, as opposed to syntagmatic similarity, i.e., the extent to which words tend to co-occur in the same units of texts. 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 (i.e., *hospital*, *nurse*, *syringe*), and semantically similar verbs are rarely related in that 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.

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 5-word context window (i.e., five words to the left and five words to the right).

```
     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.

	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

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

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.[6] 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 commonly 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.

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

In the (transformed) co-occurrence matrix, each row is a word vector, which represents the distributional profile of that word. Under the assumption that semantic distance between words is a function of their distributional similarity, 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, the cosine measure is by far the most frequently used in distributional models of word meaning; 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 taken 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, which precisely reflects 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 in the same contexts, 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 (and not only strict similarity) 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 conflated into a single word sense. While this comment is in order, whether it is an actual problem for a particular application is an empirical question. The problem does obviously not arise with monosemous words, and it is sometimes not problematic to regard related and very similar senses as a single meaning (the problem 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., clearly differentiated senses balanced in frequency) should be treated with a pinch 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). However, in this study, I will simply ignore the polysemy issue, which should not be a major problem since most of the verbs I submit to distributional classification in section 5 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 that model, semantic similarity can be quantified, which open a range of useful applications for empirical studies, such as the derivation of other quantitative measures based on semantic similarity, as well as statistical testing (cf. section 5.5). It should be emphasized, however, that vector-space modeling is merely seen in this paper as providing a proxy to word meaning (and semantic

similarity between words in particular), but I remain agnostic as to whether distributional information should be considered as a representation of meaning itself.

That being said, 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 at least are good models of human behavior. Andrews et al. (2008) evaluate the relative importance of experiential (i.e., based on properties available to the senses) and distributional information for semantic representations by comparing the performance of models based on either kind of information and one based on a combination of the two. They find that a model based on both kinds of information provides more coherent results than models based on either kind and also perform 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). Their results suggest that distributional information might well be a key component of how human beings acquire and process semantic information. Hence, my 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.

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, Wächli 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 (in press) 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 I will show, 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. In the next section, I present 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", 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 efforts deployed by 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) to (10) below 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* in [9]) and it is the quality of that performance that is intensified.

(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 depend to some extent on the context. With *ride* in (9), the event is longer and involves more strain on the vehicle than usual. In (10), *sell the hell out of* means (in this case) "sell a lot". Both examples relate to the intensification of the agent's efforts mentioned previously, and so do examples (11) and (12), which also focus on the insistence of the agent to obtain a particular result.

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

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

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

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

Instances of the 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 seem to pattern with instances of the removal construction (Johnson and Goldberg 2012) conveying the meaning 'X CAUSES Y to MOVE from Z', like *He took the gun out of the holster*. However, *the hell* clearly does not behave like a direct object in examples (3) to (12); it cannot, for example, become the subject of a

passive sentence. Semantically, *the hell* is not clearly referential (even in a figurative sense) and consequently does not bear the semantic relation that the direct object argument usually does to the verb; it is rather the referent of the prepositional phrase complement that bears this relation. Hence, *the hell* should not even be treated as a noun phrase in the traditional sense; it is rather a 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*). This fact suggests that the *hell*-construction could be treated as a case of particle insertion, whereby the sequence *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 typically transitive, uses with intransitive verbs are also attested, such as *listen* in (13).

(13) I've been listening the hell out of your tape.

This suggests that the pattern cannot be derived compositionally from any other constructions in the language, and therefore forms its own generalization. The *hell*-construction actually lends itself nicely to a construction grammar analysis (Goldberg 1995; 2006), whereby the abstract meaning of intensification is directly associated with the whole phrasal pattern "V *the hell out of* NP". The *hell*-construction is similar to more "vulgar" variants in which *hell* is replaced by other taboo words (e.g., *crap*, *fuck*, *shit*), and could therefore be considered a member of a family of related constructions (Goldberg and Jackendoff 2004).

I used the Corpus of Historical American English (COHA; Davies 2010) to collect data on the diachronic development of the verb slot in the *hell*-construction. The COHA consists of about 20 million words[7] of written American English for each decade between 1810 and 2009 and is available online. The corpus is roughly balanced for genre, in that each decade contains texts of four kinds (fiction, magazines, newspapers, non-fiction) in roughly the same proportions.[8] I queried for the string "[v*] the hell out of" in the COHA, thus retrieving instances of all verbs followed by the sequence "the hell out of". I downloaded all tokens and filtered out the instances of the *hell*-construction manually, mostly ruling out locative constructions like *get the hell out of here*.

7 20 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.

8 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.

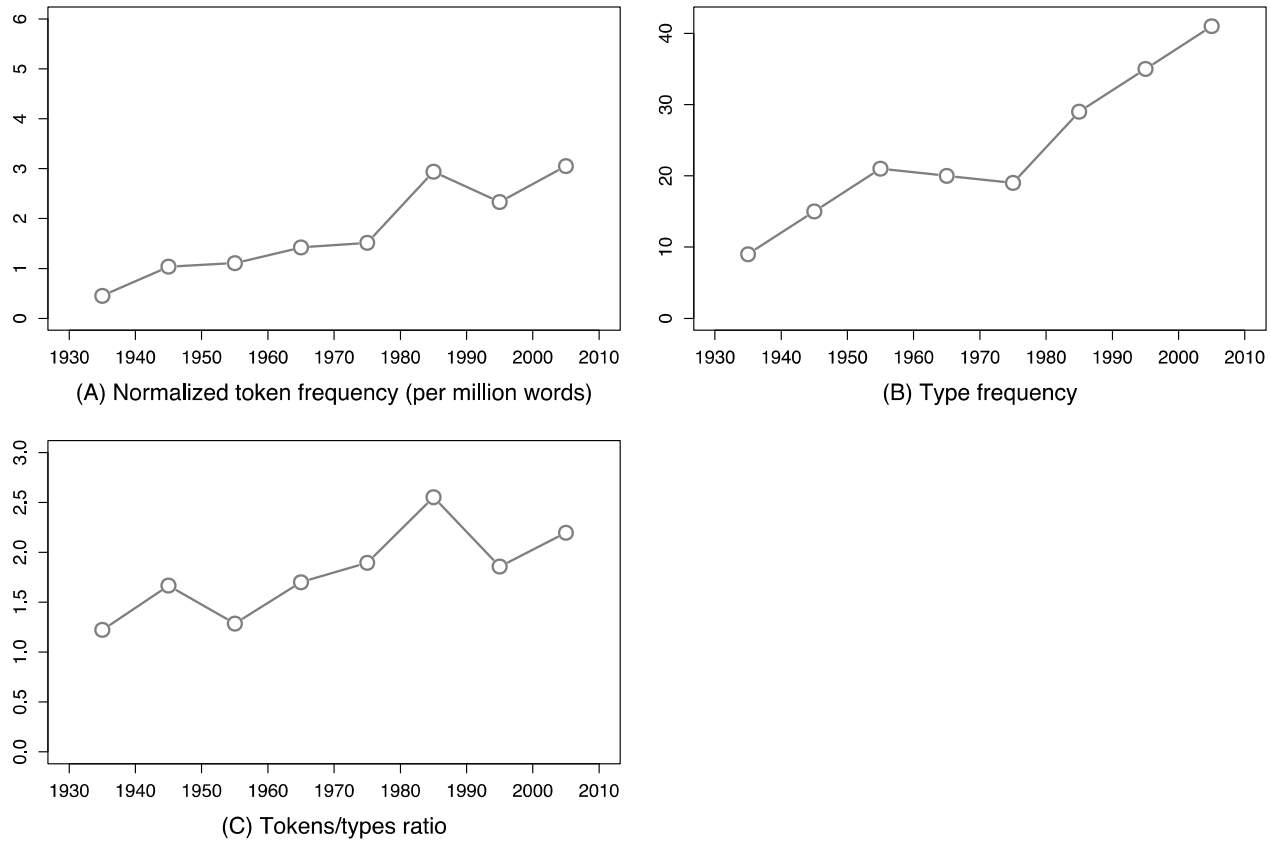

Figure 2: Diachronic development of the *hell*-construction in token frequency normalized by million words in each decade (A), type frequency (B), and token/type ratio (C) per decade.

The diachronic evolution of the verb slot in terms of token and type frequency is plotted in several diagrams in Figure 2. As can be seen from Figure 2, the first attestations of the construction in the corpus date back to the 1930s.[9] It is far beyond the province of this article to fully discuss the origin of the construction; we can only speculate that the pattern is probably related to the exclamatory uses of *(the) hell* (as mentioned above), and that it might have been derived from a figurative use of the removal construction, whereby the intensification of the action is expressed by the idea that something important (evasively referred to with the exclamatory phrase *the hell*) is forcefully removed from the patient. As 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 phenomenon, 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

9 Actually, one instance was found in 1928: "Swap generals with us and we'll lick the hell out of you" (with the verb *lick* used in the sense of 'beat, defeat'). 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.

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, I will analyze the distribution of the construction from a semantic point of view by using a measure of semantic similarity derived from distributional information. The semantic distributional model is described in section 5.2, and evaluated in 5.3. In section 5.4, I showcase two visualization techniques that use the distributional semantic similarity to investigate the diachronic development of the semantic distribution of the construction, and in section 5.5, I submit the data to statistical analysis to evaluate how the semantic structure of the distribution predicts how verbs are productively used in the construction.

5.2 The vector-space model

One of the goal 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 similarity 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 2 billion-word ukWaC corpus of blogs from the .uk domain, and Wikipedia dumps. However, I chose to use the COCA, which is also a sizable enough and well-balanced corpus of American English, and therefore 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. Although the corpus is only accessible through a web interface and cannot be downloaded in its entirety, this is not a problem for us, since the study is concerned with a limited number of verbs and not all words in the corpus.

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 proved possible to find enough data to achieve that purpose, since even the 20 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. Therefore, I am basically using data from the 1990s and 2000s to derive a measure of semantic similarity to model usage data from as early as the 1930s. This is, however, not as problematic as it might sound, since the meaning of the verbs under consideration are not likely to have changed considerably within the time frame of this study, since we are dealing with a relatively short and recent interval in which American English had long been standardized (especially in written usage) and its semantics (just like its grammar) regulated by such authoritative sources as Webster's *American Dictionary of the English Language*, whose first edition was published in 1828.[10] Besides, using the same distributional data

10 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 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

presents the advantage that we will be using the same semantic space for all time periods, which makes it easier to visualize changes.

I extracted all instances of the relevant verbs from the COCA with their context of occurrence. Verbs that I judged not frequent enough (i.e., less than 2,000 occurrences) to assess their meaning from their distribution were excluded at this stage: *bawl*, *belt*, *cream*, *dent*, *disgust*, *flog*, *grease*, *horsewhip*, *infilade*, *infuriate*, *irk*, *lam*, *micromanage*, *mortgage*, *nag*, *nuke*, *sodomize*, *squash*. This left me with 92 usable verbs. The words in the sentence contexts extracted from the COCA were lemmatized and annotated for part-of-speech using TreeTagger (Schmid 1994).[11] The matrix of co-occurrences between the target verbs and the lemma of their collocates (hence, the frequency counts of all inflected forms of a given word were collapsed into a single count) within a 5-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 nouns, verbs, adjectives, and adverbs listed among the 5,000 most frequent words in the corpus were considered as collocates (to the exclusion of the verbs *be*, *have*, and *do*),[12] 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).[13] 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).

5.3 Evaluation of the vector-space model

Before turning to the analysis of the hell-construction proper, I first evaluate the validity of the vector-space model to capture semantic similarity between verbs. To visualize similarity relations and possible groupings that can be inferred from the distributional data, I submitted the rows of the co-occurrence matrix to hierarchical clustering. 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 our 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 similarity.[14] The hierarchical clustering algorithm uses pairwise distances between rows to recursively merge the two most similar clusters (each observation is itself considered a cluster) into a higher-level cluster, until there is only one cluster containing all objects; the distance between clusters is assessed according to which linkage criterion is chosen.[15] 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 similarity between them is concerned.

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

12 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).

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

14 I used the "cosine" function from the R package lsa (Wild 2007).

15 I used the "hclust" function of the R environment.

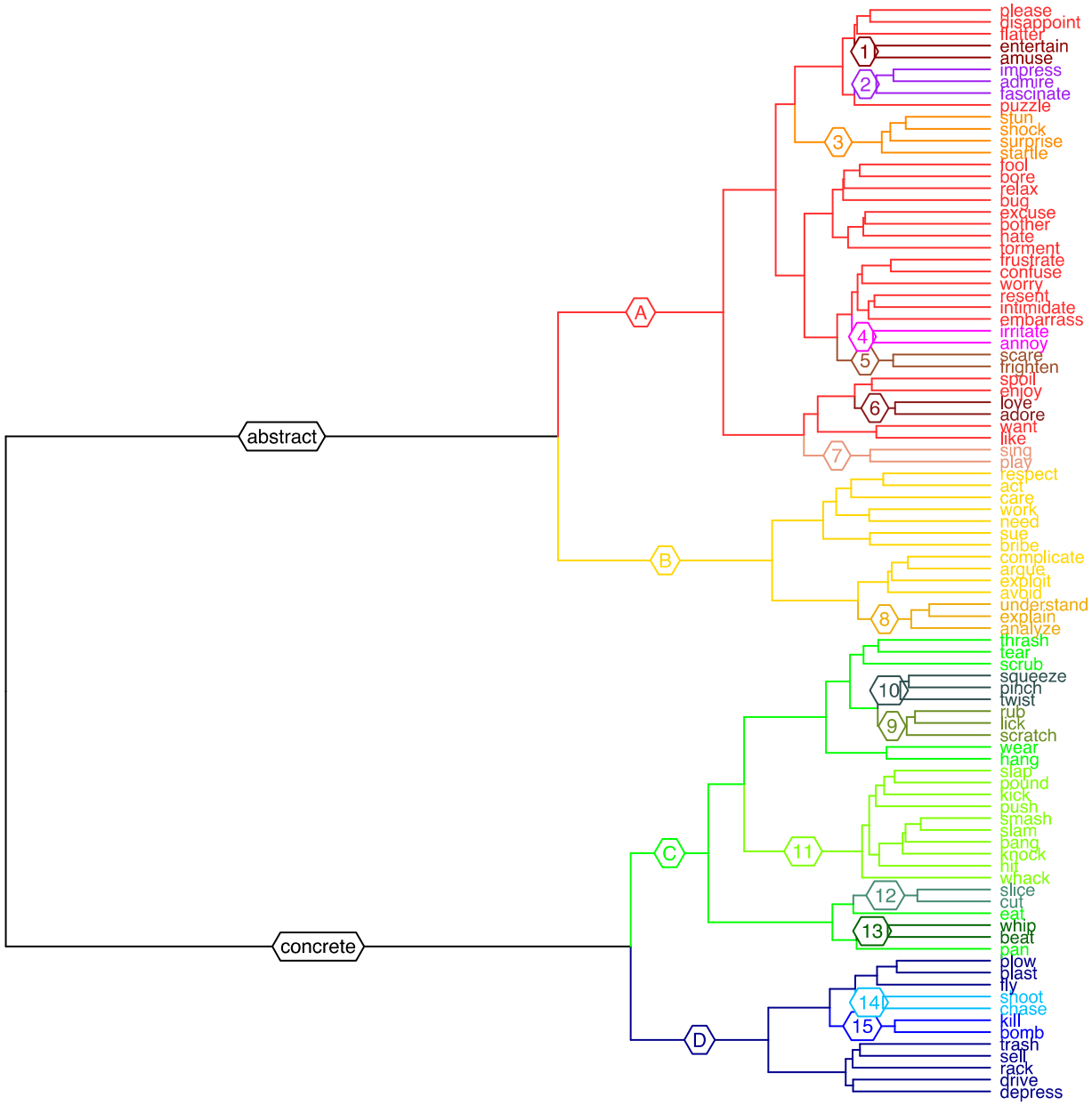

Figure 3: Cluster dendrogram for all verbs in the distribution of the *hell*-construction between 1930 and 2009.

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.[16] This kind of diagram arranges items according to their similarity, and as it were, traces the history of cluster mergers by the algorithm, from the earliest one on the right 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

16 I used Ward's linkage method to generate Figure 3. Ward's criterion aims to minimize the variance within clusters when choosing which two clusters to merge. Compared to other linkage methods, it has the property of generating more "compact" clusters.

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 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), *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* share the notion of contact and 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 violent component. This group constitutes a coherent semantic class that can evidently be derived from distributional information.

At a higher level, the clustering algorithm appears to partition the distribution in a way that 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 two semantically coherent classes each, labelled A to D in Figure 3. (A) mostly contains psych-verbs describing feelings and emotions: *please*, *surprise*, *hate*, *worry*, *annoy*, *like*, etc. (B) contains the other kinds of abstract actions. (C) mostly contains violent 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 (D), however, while they might also involve some degree of violence, typically do so less directly, and do not necessarily involve contact; a few of them are perfectly harmless actions that do not have the causative character of the verbs in group (C), e.g., *drive*, *sell*.

Much 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, however, 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 there indeed 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 very much satisfactory, which warrants its use for the study of the syntactic productivity of the *hell*-construction.

5.4 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 similarity measure can be seen as providing a way to locate words in that space with respect to each other. By aggregating the similarity 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 (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. In this section, I start the analysis of the *hell*-construction in diachrony, using two well-known visualization techniques to identify patterns in the semantic distribution.

Multidimensional scaling (MDS) provides a way both to aggregate similarity information and to represent it visually. This technique aims to place objects in a space with two (or more) dimensions such that the between-object distances are preserved as much as possible. Each object is assigned coordinates in the relevant number of dimensions by the algorithm; with two dimensions, the resulting set of coordinates can be used to generate a plot that visually depicts the similarity relations between objects.

I submitted the pairwise distances between all verbs in the distribution to multidimensional scaling into two dimensions.[17] This is essentially tantamount to mapping the high-dimensional distributional space (where each of the 4,683 collocates is one dimension) of the co-occurrence matrix into a 2-dimensional space, which, following the distributional hypothesis, should offer an approximation of the semantic space. To visualize the diachronic development of the semantic domain of the *hell*-construction, I divided the diachronic data 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. I extracted the distribution of the construction in each time period, and I plotted each distribution 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 5.3 (cf. Figure 3). For the sake of comprehensiveness, token frequency is also represented in the plots (although it will not be subsequently discussed), in that verbs with a token frequency greater than one are given a circle in addition to their label; the size of the circle is proportional to the natural logarithm of the token frequency.

17 I used the "isoMDS" function from the MASS package. The 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). This shows that the two-dimensional space returned by MDS is reasonably faithful to the original high-dimensional space.

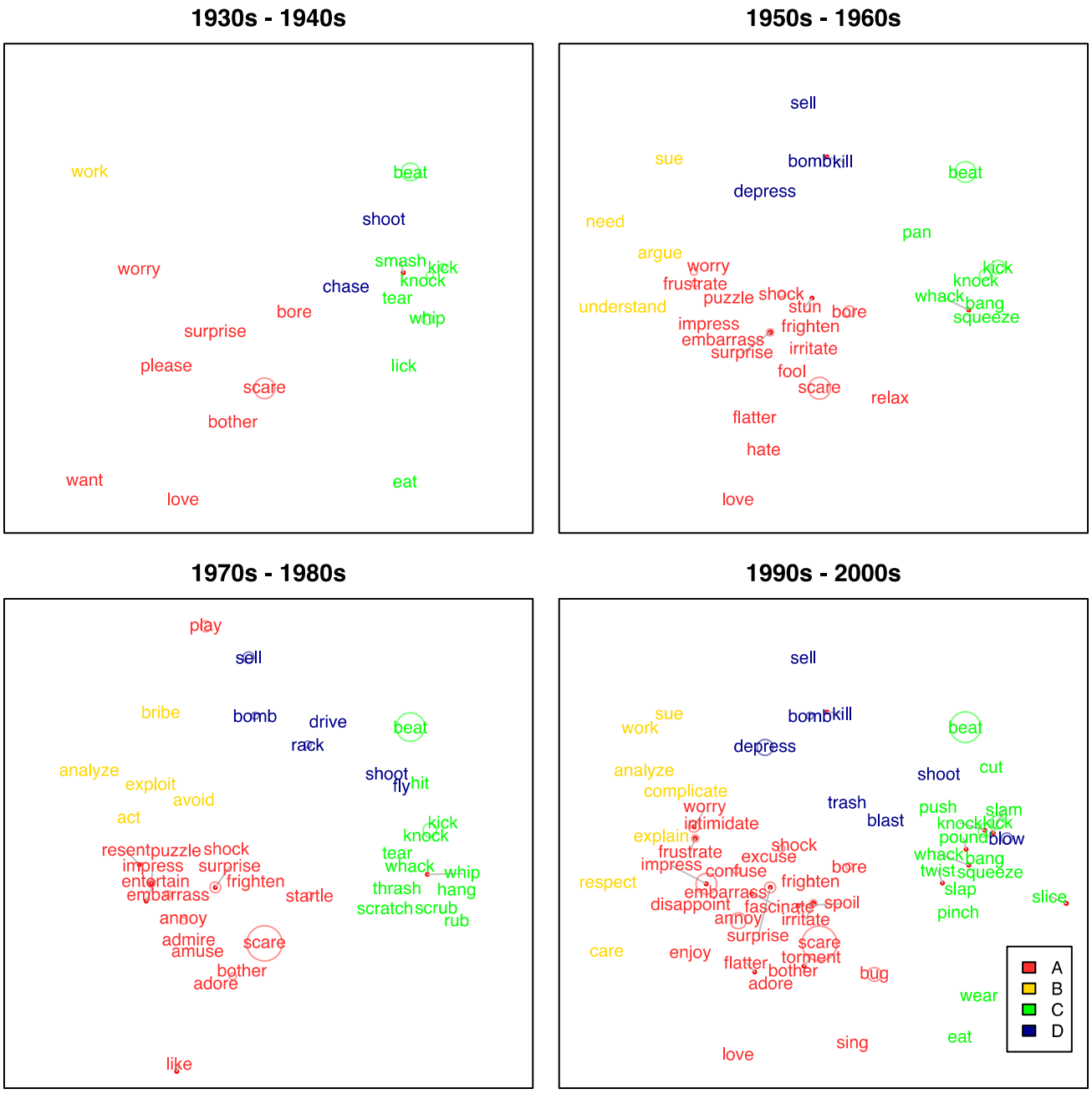

Figure 4: Semantic plots of the *hell*-construction in four successive twenty-year periods. Colors correspond to the four clusters of verbs identified by cluster analysis (cf. Figure 3).

By comparing the plots in Figure 4, we can follow the semantic development of the *hell*-construction.[18] First, one thing that is particularly striking is that the construction is clearly centered around two kinds of verbs: psych-verbs (*surprise*, *please*, *scare*, etc.) and verbs of hitting (*smack*, *kick*, *whack*, etc.), a group that is orbited by other kinds of rather violent actions (such as *pinch*, *push*, and *tear*). These two types of verbs are found in the distribution of the construction at all times. They account for the lion's share of the distribution at the onset, 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 semantic 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 (cf. footnote 17), 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 pinch of salt and its results be compared with another, more reliable method.

To analyze how regions of the semantic space fill up over time, we can use groupings based on cluster analysis (as diagrammed in Figure 3), instead of delimiting groups of verbs in the semantic plots more or less arbitrarily. If we cut the dendrogram of Figure 3 at a given level, we can obtain a list of clusters of comparable internal coherence (from a purely quantitative point of view). By combining a given clustering solution with the diachronic partition of the distribution into periods created to construct the semantic plots, we can plot how the size of each cluster varies over time. 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 figured out from the dendrogram in Figure 3, but this information is not necessary in order to observe that it is always the groups containing the most members from the onset 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

18 Note that this idea and its technical implementation are similar to the concept of motion charts, recently proposed by Hilpert (2011) to visualize linguistic change on the basis of the frequency of co-occurrence of lexical and grammatical features (the occurrence of complement-taking verbs with different kinds of finite and non-finite complement clauses). 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 (corresponding to items moving around in the plotted space), but how the semantic domain of a construction is filled at different points of its history.

of granularity; unsurprisingly, they correspond more or less to the two semantic classes described above (psych-verbs and violent actions). In sum, the results of cluster analysis are largely in line with those from the semantic plots, confirming them with a more principled and reliable method. These findings illustrate the role of type frequency in productivity: within a given semantic domain, 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.

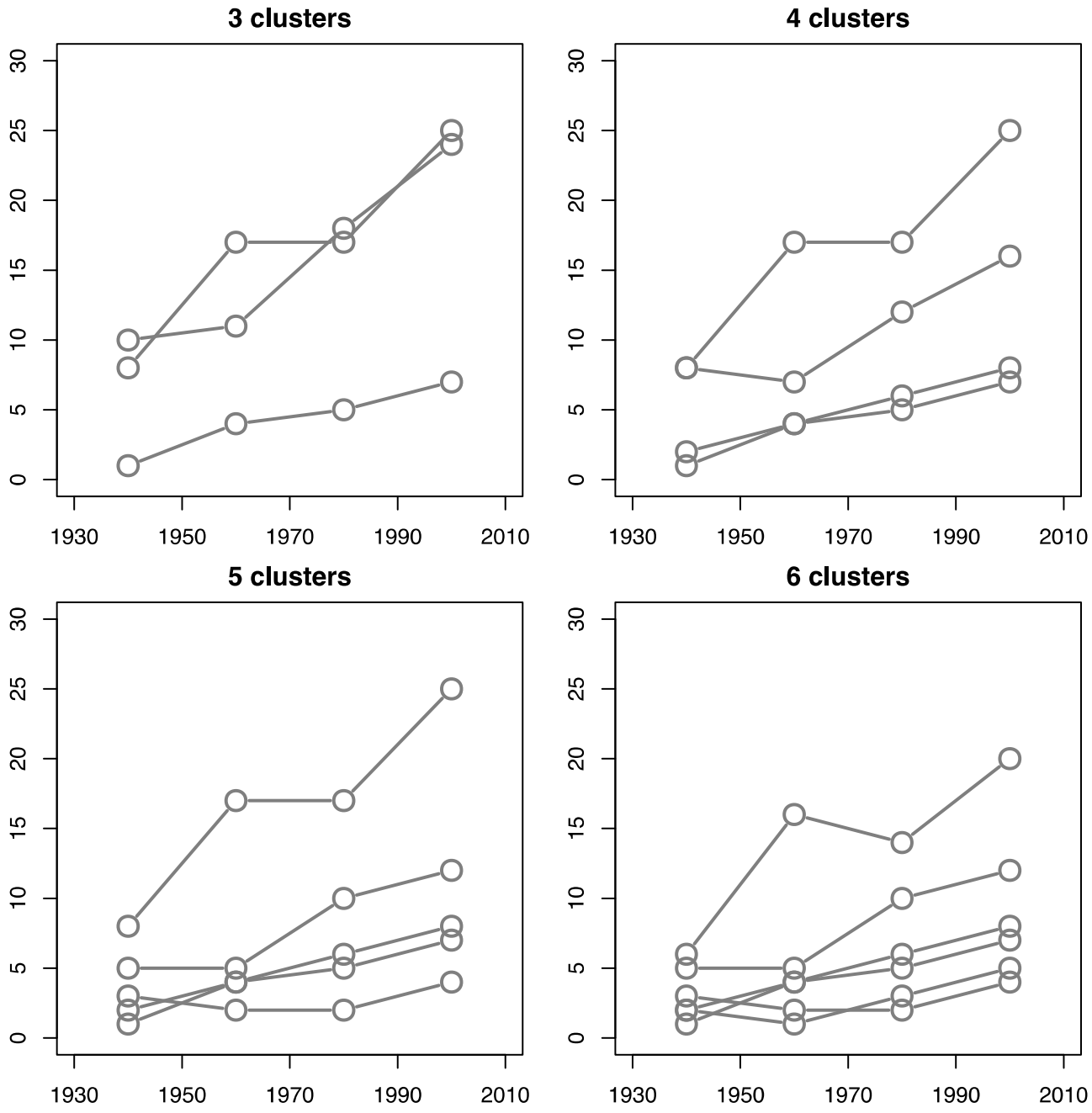

Figure 5: Type frequency variation of semantic clusters at different levels of granularity (viz., different numbers of clusters).

At the same time, there is a notable difference between verbs of violent actions 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 violent actions are rather limited, but the range of feelings and emotions experienced by human beings is much 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 which accounts for why psych-verbs turn out to be more productive than violent 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).

5.5 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 semantic aspects of a word distribution to be derived and submitted to statistical analysis. In this section, I show 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 basically the same meaning since its inception, all verbs ever attested in the construction (and probably others) 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 coinage to occur depends on properties of the prior usage of the construction (cf. section 2), especially as it relates to the presence of many 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. To test this prediction, I tried to determine 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 around that verb in the immediately preceding period. In particular, I suggest that the occurrence of a new item in the construction is related to the density of the semantic space around that item. This notion of density can be seen as an operationalization of the concept of coverage put forward by Suttle and Goldberg (2011) to explain their experimental results (cf. section 2).

For each verb in the distribution, I determined in which period the verb first occurred in the construction. For the verbs first occurring in 1970-1989 and 1990-2009,[19] I coded the binary variable OCCURRENCE, which was set as true for the first period of occurrence, and as false for all earlier periods (later periods were ignored). For each verb-period pair thus obtained, a measure of density was computed that capture 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,

19 The other verbs 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.

and the measure of density is computed on the semantic space from the third period (1970-1989). Two other datapoints 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 that need to be addressed is how to measure the density of the semantic space at a given point of that 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, a good measure of density should be defined locally, i.e., it should consider only a limited portion of the semantic space (and by no means not all of it), otherwise it will invariably be sensitive with the number of items in the space, regardless of how relevant these items are to a new coinage.

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 is equal to one minus the mean distance of the N neighbors to this item. The mean distance to nearest neighbors decreases with space density (i.e., if there are many close neighbors), and is therefore technically a measure of sparsity; since cosine distances are between 0 and 1, subtracting the mean distance from one returns a measure of density. 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 for each verb (the N variable in the formula) between 3 and 8. The predictive power of each version of the density measure was tested in a different linear model. The results of these models are summarized in Table 2 below.

Number of 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

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)

For all values of N, we find a positive effect of DENSITY, i.e., 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 receives a straightforward explanation (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). However, the variation in *p*-value is more puzzling; it basically 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 better prediction than a measure supported by fewer types. This means that productivity not only hinges on whether the existing semantic space covers the novel item, it also occurs more reliably when coverage is attested by more items. These finding support the view that semantic coverage and type frequency, while they both positively influence syntactic productivity, do so in different ways: semantic coverage sets the necessary conditions for a new coinage to occur, while type frequency increases the confidence that this coinage is indeed possible.

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 heavily 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, I consider a third alternative that avoids the limitations of both kinds of approaches, which consists 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 that it is 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, I showed how the distributional semantic approach to semantic similarity can be applied to the study of syntactic productivity in diachrony. I used multidimensional scaling and cluster analysis as means of visually representing the distributional semantic information provided by a vector-space model. I showed 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-years periods from 1930 to 2009. The results of this exploratory analysis were largely in line with current views on the determinants of syntactic productivity. Finally, I submitted the data to statistical analysis. Using mixed effects logistic regression, I found 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 constitutes empirical evidence for the relevance of Suttle and Goldberg's (2011) notion of coverage to diachronic data. Moreover, I also found that the robustness of this effect increases with the number of items that are considered in the calculation of the density measure. I interpreted this finding as illustrating the 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 similarity, it also allows for methods to be used for which quantification is necessary, such as data visualization and statistical analysis. That being said, I have but 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 to test the influence of different aspects of the semantic space (beyond density) on productivity, that the case of the *hell*-construction did not exemplify, like for instance 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, B., J. Close, G. Leech and S. Wallis, eds. (2013). *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge University Press.

Andrews, M., G. Vigliocco. and D. Vinson (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. Psychological Review 116(3), 463-498.

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

Boussidan, A. (2013). *Dynamics of semantic change: Detecting, analyzing and modeling semantic change in corpus in short diachrony*. PhD thesis, Université Lumière Lyon 2.

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

Baayen, H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.

Baayen, H., D. Davidson, and D. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390—412.

Baayen, H. and R. Lieber (1991). Productivity and English derivation: A corpus-based study. *Linguistics* 29, 801–844.

Bates, D., M. Maechler and B. Bolker (2011). *lme4: Linear mixed-effects models using S4 classes*. R package. URL: http://CRAN.R-project.org/package=lme4

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam/Philadelphia: John Benjamins.

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

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

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

Bybee, J. and S. Thompson (1997). Three frequency effects in syntax. *Berkeley Linguistics Society* 23, 65–85.

Chomsky, N. (1986). *Knowledge of language*. Cambridge, MA: MIT Press.

Cook, P. and S. 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, W. (2010). Relativity, linguistic variation and language universals. *CogniTextes* 4. URL: http://cognitextes.revues.org/303

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

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

Davies, M. (2011). *Google Books Corpus. (Based on Google Books n-grams)*. Available online at http://googlebooks.byu.edu/

Dinu, G., N. Pham and M. 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)*. East Stroudsburg PA: ACL, 31-36.

Divjak, D. and S. Gries (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1), 23-60.

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

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society.

Glenberg, A. and D. 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, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, A. and R. Jackendoff (2004). The English Resultative as a Family of Constructions. *Language* 80(3), 532-568.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer.

Gries, S. and A. Stefanowitsch (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 73-90. Stanford, CA: CSLI.

Gulordava, K. and M. 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*. East Stroudsburg, PA: ACL, 67-71.

Hilpert, M. (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, H. and B. Joseph (1996). *History, Language Change and Language Relationship. An Introduction to Historical and Comparative Linguistics*. Berlin: Walter de Gruyter.

Israel, M. (1996). The way constructions grow. In A. Goldberg (ed.), *Conceptual structure, discourse and language*. Stanford, CA: CSLI Publications, 217-230.

Krug, M. (2000). *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. Berlin: Mouton de Gruyter.

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

Landauer, T., P. Foltz, and D. Laham (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259-284.

Leech, G., M. Hundt, C. Mair and N. Smith (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20.1, 1-31.

Levin, B. (1993). English Verb Classes and Alternations: a preliminary investigation. University Of Chicago Press.

Levshina, N. and K. Heylen. (in press). A Radically Data-driven Construction Grammar: Experiments with Dutch causative constructions. In R. Boogaart, T. Colleman & G. Rutten (Eds.), *Constructions in Germanic – Extending the scope*.

Lund, K., Burgess, C. and R. Atchley (1995). Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings (LEA)*, 660-665.

Mair, C. (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, C. (2006). *Twentieth-Century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.

Miller, G. and W. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1-28.

Padó, S. and M. Lapata (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 33(2), 161-199.

Pedersen, T. (2006). Unsupervised corpus-based methods for WSD. In E. Agirre and P. Edmonds (eds.), *Word Sense Disambiguation: Algorithms and Applications*. Springer, 133–166.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press/Bradford Books.

Pantel, P. and D. 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, H. (1998). Automatic word sense discrimination. *Computational Linguistics* 24(1), 97–124.

Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.

Purandare, A. and T. 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*, pp. 41-48.

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

Ross, J. (1973). Nouniness. In O. Fujimura (ed.), *Three Dimensions of Linguistic Research*. TEC Company Ltd.

Sagi, E., S. Kaufmann and B. 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, M. (2008). The distributional hypothesis. *Rivista di Linguistica* 20(1), 33-53.

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

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

Turney, P. and P. Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Vyas, V., & Pantel, P. (2009). Semi-automatic entity set refinement. In *Proceedings of NAACL-09, Boulder, CO*.

Wälchli, B. and M. Cysouw (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3), 671-710.

Wild, F. (2007). An LSA package for R. In *Mini-Proceedings of the 1st European Workshop on Latent Semantic Analysis in Technology-Enhanced Learning, Heerlen, NL*.

Wonnacott, E., J. Boyd, J. Thompson and A. Goldberg (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language* 66, 458–478.

Zeschel, A. (2012). *Incipient productivity. A construction-based approach to linguistic creativity*. Berlin/New York: de Gruyter.