The distributional hypothesis *

Magnus Sahlgren

Distributional approaches to meaning acquisition utilize distributional properties of linguistic entities as the building blocks of semantics. In doing so, they rely fundamentally on a set of assumptions about the nature of language and meaning referred to as "the distributional hypothesis". The main point of this hypothesis is that there is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter. However, it is neither clear what kind of distributional properties we should look for, nor in what sense it is meaning that is conveyed by distributional patterns.

This paper examines these two questions, and shows that distributional approaches to meaning acquisition are rooted, and thrive, in structuralist soil. Recognizing this fact enables us to see both the potentials and the boundaries of distributional models, and above all, it provides a clear and concise answer to the above-posed questions: a distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words.

The paper discusses the structuralist origins of the distributional methodology, and distinguishes the two main types of distributional models - the syntagmatic and the paradigmatic types. It also takes a summary look at how these models are implemented, and discusses their main parameters from a linguistic point of view. The paper argues that - under the assumptions made by the distributional paradigm - the distributional representations do constitute full-blown accounts of linguistic meaning.

1. Introduction

Distributional approaches to meaning acquisition utilize distributional properties of linguistic entities as the building blocks of semantics. In doing so, they rely fundamentally on a set of assumptions about the nature of language and meaning referred to as *the distributional hypothesis*. This hypothesis is often stated in terms like "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough 1965); "words with similar meanings will

Rivista di Linguistica 20.1 (2008), pp. 33-53 *(ricevuto nell'ottobre 2008)*

occur with similar neighbors if enough text material is available" (Schütze & Pedersen 1995); "a representation that captures much of how words are used in natural context will capture much of what we mean by meaning" (Landauer & Dumais 1997); and "words that occur in the same contexts tend to have similar meanings" (Pantel 2005), just to quote a few representative examples. The general idea behind the distributional hypothesis seems clear enough: there is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter.

However, one can pose two very basic questions concerning the distributional hypothesis. The first is what kind of distributional properties we should look for, and what – if any – the differences are between different kinds of distributional properties. Looking at algorithms for distributional meaning acquisition we can discern two distinct approaches. The first is to build distributional profiles for words based on which other words surround them, as exemplified by Schütze (Schütze 1992) and the Hyperspace Analogue to Language (HAL) model (Lund et al. 1995). The second is to build distributional profiles based on in which text regions words occur, as exemplified by the Latent Semantic Analysis (LSA) model (Landauer & Dumais 1997). These approaches are often treated as functionally equivalent when it comes to representing meaning similarities, despite the fact that they are based on different types of distributional raw materials.

The second question is in what sense it is *meaning* that is conveyed by distributional patterns. Proponents of distributional methods often seem comfortable to ascribe meaning to distributional representations without explaining what they *mean* by meaning. For the non-distributionalist, on the other hand, this will surely seem odd if not completely outrageous, since meaning is usually taken to involve both reference to objects and situations in the world outside language, and to concepts and ideas inside the mind of the language user. Furthermore, if different distributional models use different types of information to extract similarities, should we not expect that they extract different *types* of similarities? And if the similarities are semantic in nature, then does it not follow that the two different approaches mentioned above should acquire different types of meaning representations?

The purpose of this paper is to examine these two questions, and to show that distributional approaches to meaning acquisition are rooted, and thrive, in structuralist soil. Recognizing this fact enables us to see both the potentials and the boundaries of distributional models, and above all, it provides a clear and concise answer to the above-posed questions.

In the following sections, we will discuss the origins of the distributional methodology, and see that it is based on structuralist assumptions about language and meaning. This will help us distinguish the two main types of distributional models, and it will also help us characterize the kind of semantic information they acquire. We will then take a summary look at how these models are implemented, and discuss their main parameters from a linguistic point of view. The last section will argue that – under the assumptions made by the distributional paradigm – the distributional representations *do* constitute full-blown accounts of linguistic meaning.

2. The distributional methodology

The distributional hypothesis is often motivated by referring to the works of Zellig Harris, who advocated a distributional methodology for linguistics. In this section, we shall see if a closer reading of Harris' ideas can help clarify the questions we identified in the introduction.

In the distributional methodology the *explanans* takes the form of distributional facts that establishes the basic entities of language and the (distributional) relations between them. Harris' idea was that the members of the basic classes of these entities behave distributionally similarly, and therefore can be grouped according to their distributional behavior. As an example, if we discover that two linguistic entities w_1, and w_2, tend to have similar distributional properties, for example that they occur with the same other entity w_3, then we may posit the *explanandum* that w_1 and w_2 belong to the same linguistic class. Harris believed that it is possible to typologize the whole of language with respect to distributional behavior, and that such distributional accounts of linguistic phenomena are "complete without intrusion of other features such as history or meaning." (Harris 1970) [1]

How does meaning fit into the distributional paradigm? Reviewers of Harris' work are not entirely unanimous regarding the role of meaning in the distributional methodology (Nevin 1993). On the contrary, this seems to be one of the main sources of controversy among Harris' commentators – how does the distributional methodology relate to considerations on meaning? On the one hand, Harris explicitly shunned the concept of meaning as part of the explanans of linguistic theory:

> As Leonard Bloomfield pointed out, it frequently happens that when we do not rest with the explanation that something is due to mean-

> ing, we discover that it has a formal regularity or 'explanation.' (Harris 1970: 785)

On the other hand, he shared with Bloomfield a profound interest in linguistic meaning; just as Bloomfield had done, Harris too realized that meaning in all its social manifestations is far beyond the reach of linguistic theory [2]. Even so, Harris was confident that his distributional methodology would be complete with regards to linguistic phenomena. The above quote continues:

> It may still be 'due to meaning' in one sense, but it accords with a distributional regularity.

What Harris is saying here is that even if extralinguistic factors *do* influence linguistic events, there will always be a distributional correlate to the event that will suffice as explanatory principle. Harris was deeply concerned with linguistic methodology, and he believed that linguistics as a science should (and, indeed, could) only deal with what is *internal* to language; whatever is in language is subject to linguistic analysis, which for Harris meant *distributional* analysis. This view implies that, in the sense that meaning is linguistic (i.e. has a purely linguistic aspect), it *must* be susceptible to distributional analysis:

> …the linguistic meanings which the structure carries can only be due to the relations in which the elements of the structure take part (Harris 1968: 2)

The distributional view on meaning is expressed in a number of passages throughout Harris' works. The most conspicuous examples are *Mathematical Structures of Language* (p. 12), where he talks about meaning being related to the combinatorial restrictions of linguistic entities; and "Distributional Structure" (p. 786), where he talks about the correspondence between difference of meaning and difference of distribution. The consistent core idea in these passages is that linguistic meaning is inherently differential, and not referential (since that would require an extra-linguistic component); it is *differences* of meaning that are mediated by *differences* of distribution. Thus, the distributional methodology allows us to quantify the amount of meaning difference between linguistic entities; it is the *discovery procedure* by which we can establish semantic similarity between words [3]:

> …if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris 1970: 786)

3. A caveat about semantic similarity

As we saw in the previous section, the distributional methodology is only concerned with meaning differences, or, expressed in different terms, with *semantic similarity*. However, as Padó & Lapata (2003) note, the notion of semantic similarity is an easy target for criticism against distributional approaches, since it encompasses a wide range of different semantic relations, like synonymy, antonymy, hyponymy, etc. Thus, it may seems as if the concept of semantic similarity is too broad to be useful, and that it is a liability that simple distributional models cannot distinguish between, e.g., synonyms and antonyms.

Such criticism is arguably valid from a prescriptive perspective where these relations are a priorily given as part of the linguistic ontology. From a descriptive perspective, however, these relations are not axiomatic, and the broad notion of semantic similarity seems perfectly plausible. There are studies that demonstrate the psychological reality of the concept of semantic similarity. For example, Miller & Charles (1991) point out that people instinctively make judgments about semantic similarity when asked to do so, without the need for further explanations of the concept; people appear to instinctively understand what semantic similarity is, and they make their judgments quickly and without difficulties. Several researchers report high inter-subject agreement when asking a number of test subjects to provide semantic similarity ratings for a given number of word pairs (Rubenstein & Goodenough 1965; Henley 1969; Miller & Charles 1991).

The point here is that the inability to further qualify the nature of the similarities in distributional models is a consequence of using the distributional methodology as discovery procedure. The distributional hypothesis, as motivated by the works of Zellig Harris, is a strong methodological claim with a weak semantic foundation. It states that *differences* of meaning correlate with *differences* of distribution, but it neither specifies *what kind* of distributional information we should look for, nor *what kind* of meaning differences it mediates. This does not necessarily mean that the use of the distributional methodology as discovery procedure in distributional models is not

well motivated by Harris' distributional approach. On the contrary, it is; but if we want to uncover the *nature* of the differences, we need to thoroughly understand the differential view on meaning.

4. The origin of differences

The differential view on meaning that Harris assumes in his distributional methodology does not originate in his theories. Rather, it is a consequence of his theoretical ancestry. Although Harris' primary source of inspiration was Bloomfield, the origin of the differential view on meaning goes back even further, to the cradle of structuralism and the *Cours de linguistique générale* (1916/1983). It is in this work Ferdinand de Saussure lays the foundation for what will later develop into structuralism.

As the word suggests, the structuralist is primarily interested in the *structure* of language, and less so in individual *usage* of it. The reason is that the abstract principles of language as a system – referred to as *la langue* – are constitutive for any individual utterance – referred to as *parole*. Saussure illustrated the idea using chess as an analogy. Chess is defined by the rules of the game together with the pieces and the board. Individual moves and actual games of chess are only interesting to the participants, and are not essential to (and may even obscure) the definition of the game. In the same manner, individual utterances are certainly interesting to the language users, but are not essential for (and may even obscure) the description of the language system.

To continue the chess analogy, the individual pieces of the game are identified by their functional differences; the king moves one step at a time in any direction, while bishops move diagonally as many steps as desired. Similarly in la langue, signs are identified by their functional differences. Saussure used the term *valeur* to describe the function of a sign. This is arguably the most important concept in structuralist theory, since it is a sign's valeur that defines its role within the language system. Valeurs are defined purely differentially, so that a sign has a valeur only by virtue of being different from the other signs. Such a differential view on the functional distinctiveness of linguistic elements highlights the importance of the system as a whole, since differences (i.e. valeurs) cannot exist in isolation from the system itself. A single isolated sign cannot enter into difference relations, since there are no other signs to differ from. In this view, the system itself becomes an interplay of functional differences:

> In the language itself, there are only differences. (Saussure 1916/1983: 166-118)

The concept of *valeur* corresponds to the idea of a thoroughly linguistic aspect of meaning. Consider the difference between the French word *mouton* and the English word *sheep*. These words may be said to have the same extralinguistic (i.e. referential) meaning, but they do not have the same valeur, since English makes a distinction between *mutton* and *sheep* that is not available in French. Thus, the functional differences between the signs within la langue is the key to the idea of linguistic meaning, and Saussure divides these functional differences into two kinds: *syntagmatic* and *paradigmatic* relations.

Syntagmatic relations concern positioning, and relate entities that co-occur in the text; it is a relation *in praesentia*. This relation is a linear one, and applies to linguistic entities that occur in sequential combinations. One example is words that occur in a sequence, as in a normal sentence like *the wolf is hungry*. Syntagmatic relations are combinatorial relations, which means that words that enter into such relations can be combined with each other. A syntagm is such an ordered combination of linguistic entities. For example, written words are syntagms of letters, sentences are syntagms of words, and paragraphs are syntagms of sentences.

Paradigmatic relations concern substitution, and relate entities that do not co-occur in the text; it is a relation *in absentia*. Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words *hungry* and *thirsty* in the sentence *the wolf is* [*hungry*|*thirsty*]. Paradigmatic relations are substitutional relations, which means that linguistic entities have a paradigmatic relation when the choice of one excludes the choice of another. A paradigm is thus a set of such substitutable entities.

The syntagmatic and paradigmatic relations are usually depicted as orthogonal axes in a 2-dimensional grid:

	Paradigmatic relations Selections: "*x* or *y* or…"			
Syntagmatic relations Combinations: "*x* and *y* and…"	she he they	adores likes love	green blue red	paint dye colour

The Saussurian notion of valeur as functional difference along the syntagmatic and paradigmatic axes is the origin of the differential view on meaning so prevalent in structuralist theories. Although Harris was arguably more directly influenced by the works of Bloomfield than of Saussure, the latter's structuralist legacy is foundational for both Bloomfield's and Harris' theories, and the differential view on meaning is decidedly foundational for the distributional hypothesis. Armed with this new-found theoretical insight and terminology, we may answer the questions posed in the introduction: *what kind* of distributional information should we look for, and *what kind* of meaning differences does it mediate?

A Saussurian refinement of the distributional hypothesis not only clarifies the semantic pretensions of distributional approaches to meaning acquisition, but it also elucidates the distributional methodology in itself. As we have seen in this section, words have a syntagmatic relation if they co-occur, and a paradigmatic relation if they share neighbors. Thus, we should be able to populate a distributional model with syntagmatic relations if we collect information about which words tend to co-occur, and with paradigmatic relations if we collect information about which words tend to share neighbors. Instead of talking about unqualified semantic similarities mediated by unspecified distributional patterns, we can now state concisely that:

> THE REFINED DISTRIBUTIONAL HYPOTHESIS: A distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words.

5. Syntagmatic models

As we saw in the previous section, a syntagmatic relation holds between words that co-occur. The prime example of co-occurrence events is collocations, such as *hermetically sealed*, where the first part *hermetically* very seldom occurs without the second part *sealed*. Collocations are probably the most obvious examples of syntagmatically related words, because the parts of the collocation tend to occur next to each other, without any intervening words. However, syntagmatically related words can also be defined as words that co-occur

within the same text region, with a (possibly large) number of words between them. In the same sense as "distributional" and "approaches" constitute a syntagmatically related word pair in a number of places throughout this paper, we could say that any two words in this paragraph (or section, or paper, or even the entire journal) constitute a syntagmatically related word pair. Thus, there is at least one parameter that applies to syntagmatic models:

1. The size of the context region within which co-occurrences are counted.

Distributional models tend to favor the use of larger text regions as context. The reason for this seems to be primarily that syntagmatic distributional approaches hail from the information-retrieval community, where a document is a natural context of a word. To see why, consider the information-retrieval universe, in which *documents* and *words* are two of the most basic elements. Documents are assumed to represent topical units (and consequently also topical *unities*), whereas words are seen as topic indicators, whose distribution is governed by a limited number of topics. In the standard type of information retrieval, this is as far as the metaphor goes, and elements of the universe (e.g. queries and documents) are matched based on word overlap, without utilizing the topics. In more sophisticated topic-based information retrieval such as LSI (Deerwester et al. 1990), the topics constitute the fundamental ontology, and all elements in the universe – such as words and documents – can be grouped according to them. In that way, queries and documents can be matched according to their topicality, without necessarily having to share vocabulary. Note that in both types of information retrieval, documents constitute the natural context of words.

This is a perfectly feasible simplification of textual reality when viewed from an information-retrieval perspective. However, information retrieval is an artificial problem, and a "document" in the sense of a topical unit–unity is an artificial notion that hardly exists elsewhere; before the advent of library science, the idea that the content of a text could be expressed with a few index terms must have seemed more or less appalling. In the "real" world, content is something we reason about, associate to, and compare. The uncritical assimilation of the information-retrieval community's conception of context is unfortunate, since the simplification is uncalled for, and may even be harmful, outside the information-retrieval universe. In the world beyond information-retrieval test collections (which tend to consist of

text types for which the metaphor actually makes sense, such as short newswire articles or downloaded web pages), text (and, in the big picture, language) is a continuous flow where topics intertwine and overlap. In this complex structure, finding a correlate to the information-retrieval notion of a *document* is at best an arbitrary choice. As Ruge (1992) notes:

> Inside of a large context (e.g. a whole document) there are lots of terms not semantically compatible. In large contexts nearly every term can co-occur with every other; thus this must not mean anything for their semantic properties. (p. 318)

So what would be a more linguistically justified definition of context in which to collect syntagmatic information? Perhaps a clause or a sentence, since they seem to be linguistic universals; clauses and sentences, or at least the functional equivalent to such entities (i.e. some sequence delimited by some kind of delimiter), seem to exist in every language – spoken as well as written or signalled. Thus, it would be possible to argue for its apparent linguistic reality as context. Sentences have been used to harvest distributional information by, e.g., Rubenstein & Goodenough (1965), Miller & Charles (1991), and Leacock et al. (1996).

Another possibility would be to use a smaller context region consisting of only a couple of consecutive words, as in the example with collocations. However, a serious problem with using such a small context to collect syntagmatic information is that very few words – basically only collocations – co-occur often within a small context region. In fact, as, e.g., Picard (1999) points out, the majority of terms never co-occur. The smaller the context regions are that we use to collect syntagmatic information, the poorer the statistical foundation will be, and consequently the worse the sparse-data problem will be for the resulting representation.

Regardless of whether one favors the use of documents, sentences, or phrases for harvesting co-occurrences, the implementational basis is the same; syntagmatic models collect text data in a words-by-documents co-occurrence matrix in which the cells indicate the (normalized) frequency of occurrence of a word in a document (or, as we have discussed in this section, some other type of text region). Table 1 demonstrates the idea w_1; has occurred one time in document 2, while w_2 has occurred one time in document 3 and three times in document 6. The point of this representation is that we can compare the row vectors – called *context vectors* – using linear algebra, so that

words that have occurred in the same documents will get high pairwise similarity scores. In this example, w_3 and w_4 have similar co-occurrence profiles, and get a score of 0.71 [4], indicating that they have occurred syntagmatically in this particular (fictive) data.

Word	Documents							
	1	2	3	4	5	6	7	8
w_1	0	1	0	0	0	0	0	0
w_2	0	0	1	0	0	3	0	0
w_3	1	0	0	2	0	0	5	0
w_4	3	0	0	1	1	0	2	0
w_5	0	1	3	0	1	2	1	0
w_6	1	2	0	0	0	0	1	0
w_7	0	1	0	1	0	1	0	1
w_8	0	0	0	0	0	7	0	0

Table 1: Words-by-documents co-occurrence matrix.

Figure 1 demonstrates an example of a syntagmatic neighborhood for the word *knife*. This particular distributional model was built from a ten-million word balanced corpus of English high-school level texts, using sections spanning approximately 150 words as contexts. Noni and Nimuk are the names of a boy and his dog in a story where a knife plays a key role.

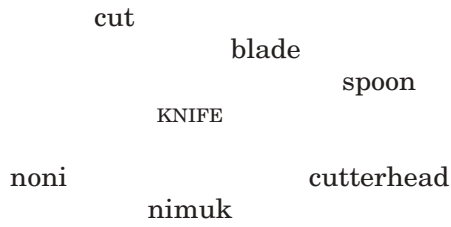

Figure 1: Syntagmatic neighborhood of *knife*.

6. Paradigmatic models

Turning now to the paradigmatic models, we begin by recalling from Section 4 that paradigmatically related words are words that do not themselves co-occur, but whose surrounding words are often the same. One example of such paradigmatically related words is dif-

ferent adjectives that modify the same nouns – e.g. *bad* and *good* in *bad news*, *good news*. As with syntagmatically related words, paradigmatic relations need not only consist of words that share the same immediately preceding or succeeding neighbor or neighbors. The paradigmatic relation may just as well be defined as words that share some, or several, of the nearest preceding or succeeding neighbors. Thus, there are at least three parameters that apply to paradigmatic models:

1. The size of the context region within which paradigmatic information is collected.
2. The position of the words within the context region.
3. The direction in which the context region is extended (preceding or succeeding neighbors).

Paradigmatic models collect distributional information using a context window of some size and extension. As an example, imagine the following two imaginary word sequences:

to have a splendid time in Rome
to have a wonderful time in Rome

Notice that *splendid* and *wonderful* constitute a paradigmatically related word pair in this example, and that it would suffice to look at the immediately preceding and succeeding words to establish this – what we call a 1+1-sized context window. In the same manner, a 2+2-sized context window would consist of the two preceding and the two succeeding words, and a 5+3-sized window of the five preceding and the three succeeding words, and so on. Naturally, nothing precludes us from using an entire sentence (or, for that matter, an entire text) as context window, but – as we will soon see – most researches favor the use of statically-sized context windows. The context window is normally advanced one word at a time until the entire data has been processed – a so-called *sliding* context window.

As noted in the previous section, we also need to account for the position of the words within the context windows, since paradigmatically related words may also be defined as words that share some of the *s* surrounding words. For example, imagine that the word *splendid* in the example above could take any of a number of different modifiers, so that one sequence would be arbitrarily realized as *a really splendid time*, and the other as *a particularly wonderful time*. In this case, we would like to exclude the modifiers from the context

windows. This can easily be accomplished by using a null-weight for that position in the context window, so that the configuration for, e.g., a 1+2-sized window would be 1 + 0 1, where 0 means that the word is ignored. This would then be realized in the example as:

> a really splendid time → splendid: (a 0) + (time)
> a particularly wonderful time → wonderful: (a 0) + (time)

The million-Euro question regarding context windows is their size: how many words to the left and to the right should we count? There have been many suggestions in the literature. For example, Schütze (1992) uses a window size of 1000 *characters*, with the argument that a few long words are possibly better than many short words, which tend to be high-frequency function words. Yarowsky (1992) uses 100 words, while Gale et al. (1994) uses 50 words to the left and 50 words to the right, with the argument that this kind of large context is useful for "broad topic classification". Schütze (1998) uses a 50-word window, whereas Schütze & Pedersen (1997) uses a context window spanning 40 words. Niwa & Nitta (1994) uses a 10+10-sized window, and the Hyperspace Analogue to Language (HAL) (1995) algorithm uses a directional 10-word window. Black et al. (1988) uses narrow windows spanning 3–6 words, Church & Hanks (1989) used 5 words, and Dagan et al. (1993) uses a window spanning 3 words, when ignoring function words.

As we can see, examples of window sizes range from 100 words to just a couple of words. There is very seldom a theoretical motivation for a particular window size. Rather, the context window is often seen as just another experimentally determinable parameter. Levy et al. (1998) is a good example of this viewpoint:

> These and other technical and practical questions can only be answered by careful and time-consuming experimentation. (p. 4 in the offprint)

Although there is undoubtedly some truth in this statement, there seems to be some empirical evidence for the feasibility of using a fairly small context window. Kaplan (1955) asked people to identify the sense of a polysemous word when they were shown only the words in its immediate vicinity. They were almost always able to determine the sense of the word when shown a string of five words – i.e. a 2+2-sized context window. This experiment has been replicated with the same result by Choueka & Lusignan (1985). Our previous experiments (Karlgren & Sahlgren 2001) also indicate that

a narrow context window is preferable to use for acquiring paradigmatic information.

As with the syntagmatic models, the implementational details of paradigmatic models are the same regardless of our choice of context window; paradigmatic models collect text data in a words-by-words co-occurrence matrix that is populated by counting how many times words occur together within the context window. Table 2 demonstrates the idea using the example sentence *whereof one cannot speak thereof one must be silent*. Note that the row and column vectors for the words are different; the row vectors contain co-occurrence counts with words that have occurred one position to the right of the words, while the column vectors contain co-occurrence counts with words that have occurred one position to their left. This type of words-by-words matrix is called a *directional* co-occurrence matrix, and it allows us to compare the right and left contexts separately, or, by concatenating the rows and column vectors for a word, the complete context profile for a word. If we instead count co-occurrences symmetrically in both directions within the context window, we end up with a *symmetric* words-by-words co-occurrence matrix in which the rows equals the columns. Note that, regardless of whether we use a directional or symmetric words-by-words matrix, we will find that words that have occurred with the same other words – i.e. that are in a paradigmatic relationship in the particular data we are currently looking at – get similar representations if we compare their context vectors.

Word	Co-occurrents							
	whereof	one	cannot	speak	thereof	must	be	silent
whereof	0	1	0	0	0	0	0	0
one	0	0	1	0	0	1	0	0
cannot	0	0	0	1	0	0	0	0
speak	0	0	0	0	1	0	0	0
thereof	0	1	0	0	0	0	0	0
must	0	0	0	0	0	0	1	0
be	0	0	0	0	0	0	0	1
silent	0	0	0	0	0	0	0	0

Table 2: Directional words-by-words co-occurrence matrix.

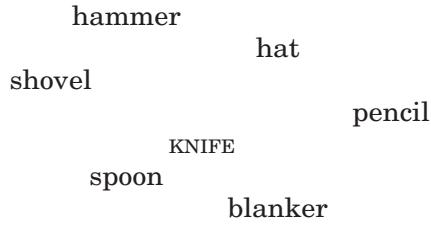

Figure 2: Paradigmatic neighborhood of *knife*.

As an example, Figure 2 shows a paradigmatic neighborhood for the word *knife*, using the same data as for Figure 1. This paradigmatic model was built using a symmetric context window spanning 2+2 words. Note that *spoon* occurs as a neighbor to *knife* in both the syntagmatic neighborhood in Figure 1 and in the paradigmatic neighborhood in Figure 2. This is symptomatic of the connection between syntagmatic and paradigmatic relations: they are not mutually exclusive, and some words occur in both syntagmatic and paradigmatic relations, to different degrees. However, experiments in Sahlgren (2006) shows that the overlap between syntagmatic and paradigmatic models is very small.

7. And what about linguistics?

Linguists that encounter distributional models tend to frown upon the lack of linguistic sophistication in their definitions and uses of context. We are, after all, throwing away basically everything we know about language when we are simply counting surface (co-)occurrences. Considering the two different types of distributional models discussed in the previous sections, the paradigmatic models are arguably more linguistically sophisticated than the syntagmatic ones, since a context window at least captures some rudimentary information about word order. But why pretend that linguistics never existed – why not use linguistic knowledge explicitly?

There have been a few attempts at using linguistic knowledge when collecting distributional information. In Karlgren & Sahlgren (2001), we used lemmatized data to increase the performance of our distributional model on a synonym test. We also experimented with adding part-of-speech tags to the words, thus performing grammatical disambiguation. However, adding part-of-speech information *decreased* the performance for all sizes of the context window, except

when using a minimal 1+1-sized window. Wiemer-Hastings & Zipitria (2001) also noticed a decrease in performance of LSA when they added part-of-speech tags to the words, and Widdows (2003) noted that adding part-of-speech information improves the representation for common nouns, but not for proper nouns or finite present-tense verbs when enriching the WordNet taxonomy. The reason for the decrease in performance is that adding part-of-speech information increases the number of unique words in the data, thus aggravating the sparse-data problem.

A more sophisticated approach to utilizing linguistic information is Padó & Lapata (2007), who uses syntactically parsed data to build contexts that reflect the dependency relations between the words. Their approach is inspired by the works of Strzalkowski (1994) and Lin (1997), who also used parsed data to compute distributional similarity between words. In Lin's experiments, words were represented by the frequency counts of all their *dependency triplets*. A dependency triplet consists of two words and the grammatical relationship between them in a sentence, such as (have subj I) from the sentence *I have angst*. Similarity between words was then defined using an information-theoretic similarity measure.

Other attempts at using linguistic information for computing distributional similarity between words include Hindle (1990), who used predicate-argument structure to determine the similarity of nouns; Hearst (1992), who extracted hyponyms using lexical–syntactic templates; Ruge (1992), who used head–modifier relations for extracting similar words; and Grefenstette (1992), who also used syntactic context to measure similarity between words.

Furthermore, recent advances in techniques for producing distributional models have made it possible to utilize full-blown word order for acquiring paradigmatic representations (Jones & Mewhort 2007; Sahlgren et al. 2008), but evidence for its usefulness remain inconclusive [5]. This seems to be a common thread in the use of linguistically refined notions of context for distributional modelling: empirical evidence for the supremacy of such refined contexts are still scarce. In addition to this, linguistically refined contexts normally require a non-negligible amount of preprocessing, and tend to suffer from sparse data (Schütze 1998). Much more research is needed in order to determine the viability of, e.g., word order or dependency relations for building distributional models of meaning acquisition.

8. Concluding thoughts

As we have seen in this paper, distributional approaches to meaning acquisition rely on a structuralist view on language, in which the only semantics available – and allowed in the model – are syntagmatic and paradigmatic relations between words. Understanding the structuralist foundation of the distributional paradigm has not only clarified the semantic pretensions of distributional methods, but has also given us tools to bar the way for the common critique raised against distributional methods that they alone are insufficient for arriving at full-blown semantics because they only look at distributional patterns in text. As this paper has demonstrated, such critique is irrelevant because the only meanings that exist within a structuralist account of language are the types of relations distributional methods acquire.

Distributional models are models of word meaning. Not the meanings that are in our heads, and not the meanings that are out there in the world, but the meanings that are in the text. The distributional paradigm might be an odd bird, but it nevertheless constitutes a viable way to meaning. Out of the plethora of theories about meaning available for the aspiring semanticist, remarkably few have proven their mettle in actual implementation. For those that have, there is usually a fair amount of fitting circles into squares going on; the theoretical prescriptions often do not fit observable linguistic data, which tend to be variable, inconsistent and vague. Semantics has been, and still is, a surprisingly impractical occupation.

In keeping with this theoretical lopsidedness, there is a long and withstanding tradition in linguistics to view the incomplete, noisy and imprecise form of natural language as an obstacle that obscures rather than elucidates meaning. It is very common in this tradition to claim that we therefore need a more exact form of representation that obliterates the ambiguity and incompleteness of natural language. Historically, logic has often been cast in this role, with the idea that it provides a more stringent and precise formalism that makes explicit the semantic information hidden in the imprecise form of natural language. Advocates of this paradigm claim that we should not model natural language use, since it is noisy and imprecise; instead, we should model language in the abstract.

In stark contrast to such a prescriptive perspective, proponents of descriptive approaches to linguistics argue that ambiguity, vagueness and incompleteness are essential properties of natural language that should be nourished and utilized; these properties are not signs

of communicative malfunction and linguistic deterioration, but of communicative prosperity and of linguistic richness. Descriptivists argue that it would be presumptuous to believe that the single most complex communication system developed in nature could be more adequately represented by some human-made formalism. Language has the form it has because it is the most viable form. In the words of Ludwig Wittgenstein (1953):

> It is clear that every sentence in our language 'is in order as it is.' That is to say, we are not *striving after* an ideal, as if our ordinary vague sentences had not yet got a quite unexceptional sense, and a perfect language awaited construction by us. (§98)

Distributional approaches to meaning acquisition are based entirely on language data, which means that they embody a thoroughly descriptive perspective. They do not rely on a priori assumptions about language (or at least they do so to a bare minimum). By grounding the representations in actual usage data, distributional approaches only represent what is *really there* in the current universe of discourse. When the data changes, the distributional model changes accordingly; if we use an entirely different set of data, we will end up with an entirely different distributional model. Distributional approaches acquire meanings *by virtue of* being based entirely on noisy, vague, ambiguous and possibly incomplete language data.

We conclude this paper with the observation that distributional models are not only grounded in empirical observation, but – as this paper as shown – they also rest on a solid theoretical foundation.

Address of the Author:
Magnus Sahlgren, SICS, Box 1263, SE16429 Kista, Sweden
<mange@sics.se>

Notes

[*] This paper is based on several chapters of my PhD dissertation (Sahlgren 2006).

[1] Harris did not exclude the possibility of other scientific studies of language: "It goes without saying that other studies of language – historical, psychological, etc. – are also possible, both in relation to distributional structure and independently of it" (Harris 1970: 775).

[2] "Though we cannot list all the co-occurrents […] of a particular morpheme, or define its meaning fully on the basis of these" (Harris 1970: 787).

[3] Note that Harris talks about meaning *differences*, but that the distributional hypothesis professes to uncover meaning *similarities*. There is no contradiction in this, since differences and similarities are, so to speak, two sides of the same coin.
[4] Using cosine similarity, which computes the angles between the vectors and normalizes for vector length. A score close to 1 indicates similarity, while a score close to 0 means they are completely unrelated.
[5] In Sahlgren et al. (2008), we show that taking account of directional information improves the quality of the paradigmatic representations, but that no improvement can be seen when using full-blown word order.

Bibliographical references

Black Ezra 1988. An experiment in computational discrimination of english word senses. *IBM J. Res. Dev.*, 32(2). 185–194.

Choueka Yaacov & Serge Lusignan 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19(3). 147–158.

Church Kenneth & Patrick Hanks 1989. Word association norms, mutual information, and lexicography. *Proceedings of the 27th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics. 76-83.

Dagan Ido, Shaul Marcus & Shaul Markovitch 1993. Contextual word similarity and estimation from sparse data. *Proceedings of the 31st Conference of the Association for Computational Linguistics*. Association for Computational Linguistics. 164-171.

Deerwester Scott, Susan Dumais, George Furnas, Thomas Landauer & Richard Harshman 1990. Indexing by latent semantic analysis. J*ournal of the Society for Information Science*, 41(6). 391–407.

Gale William, Kenneth Church & David Yarowsky 1994. Discrimination decisions for 100,000 dimensional spaces. In Zampolli Antonio, Nicoletta Calzolari & Martha Palmer (eds.). *Current issues in Computational Linguistics: In honour of Don Walker*. Dordrecht: Kluwer Academic Publishers. 429-450.

Grefenstette Gregory 1992. Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. *Proceedings of the 30th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics. 324-326.

Harris Zellig 1968. *Mathematical Structures of Language*. Interscience Publishers.

Harris Zellig 1970. Distributional structure. In *Papers in Structural and Transformational Linguistics*. pp. 775-794.

Hearst Marti 1992. Automatic acquisition of hyponyms from large text porpora. *Proceedings of the 14th International Conference on Computational Linguistics*. Association for Computational Linguistics. 539-545.

Henley Nancy 1969. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior* 8. 176-184.

Hindle Donald 1990. Noun classification from predicate-argument structures. *Proceedings of the 14th International Conference on Computational Linguistics*. Association for Computational Linguistics. 268-275.

JONES Michael & Douglas MEWHORT 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* 114(1). 1-37.

KAPLAN Abraham 1955. An experimental study of ambiguity and context. *Mechanical Translation* 2(2). 39-46.

KARLGREN Jussi & Magnus SAHLGREN 2001. From words to understanding. In UESAKAYoshinori, Pentti KANERVA & Hideki ASOH (eds.). *Foundation of Real-World Intelligence*. Stanford, California: CSLI Publications. 294-308.

LANDAUER Thomas & Susan DUMAIS 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2). 211-240.

LEACOCK Claudia, Geoffrey TOWELL & Ellen VOORHEES 1996. Towards building contextual representations of word senses using statistical models. In BOGURAEV Branimir & James PUSTEJOVSKY (eds.). *Corpus processing for lexical acquisition.* Cambridge, MA, USA: MIT Press. 97-113.

LEVY Joseph, John BULLINARIA & Malti PATEL 1998. Exploration in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology* 10(1). 99-111.

LIN Dekang 1997. Using syntactic dependency as local context to resolve word sense ambiguity. *Proceedings of the 35th International Conference of the Associations for Computational Linguistics*. 64-71.

LUND Kevin, Curt BURGESS & Ruth ATCHLEY 1995. Semantic and associative priming in high dimensional semantic space. *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. 660-665.

MILLER George & Walter CHARLES 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1). 1-28.

NEVIN Bruce 1993. A minimalist program for linguistics: The work of Zellig Harris on meaning and information. *Historiographia Linguistica* 20 (2/3). 355-398.

NIWA Yoshiki & Yoshihiko NITTA 1994. Cooccurrence vectors from corpora vs. distance vectors from dictionaries. *Proceedings of the 15th International Conference on Computational Linguistics*. Association for Computational Linguistics. 304-309

PADÓ Sebastian & Mirella LAPATA 2003. Constructing semantic space models from parsed corpora. *Proceedings of the 41st Conference of the Association for Computational Linguistics*. 128-135.

PADÓ Sebastian & Mirela LAPATA 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2). 161-199.

PANTEL Patrick 2005. Inducing ontological cooccurrence vectors. *Proceedings of the 43rd Conference of the Association for Computational Linguistics*. Association for Computational Linguistics. 125-132.

PICARD Justin 1999. Finding content-bearing terms using term similarities. Proceedings of the 9th Conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 241-244.

RUBENSTEIN Herbert & John GOODENOUGH 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10). 627-633.

RUGE Gerda 1992. Experiments on linguistically-based term associations. *Information Processing and Management* 28(3) 317-332.

SAHLGREN Magnus 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in highdimensional vector spaes*. Department of Linguistics, Stockholm University. PhD Dissertation.

SAHLGREN Magnus, Anders HOLST & Pentti KANERVA 2008. Permutations as a means to encode order in word space. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. 1300-1305.

SAUSSURE Ferdinand (1916) 1983. *Course in General Linguistics*. (Translated by Roy Harris). London: Duckworth.

SCHÜTZE Hinrich 1992. Dimensions of meaning. *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*. 787-796.

SCHÜTZE Hinrich 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97-123.

SCHÜTZE Hinrich & Jan PEDERSEN 1995. Information retrieval based on word senses. *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. 161-175.

SCHÜTZE Hinrich & Jan PEDERSEN 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management* 33(3). 307-318.

STRZALKOWSKI Tomek 1994. Building a lexical domain map from text corpora. *Proceedings of the 15th International Conference on Computational Linguistics*. 604-610.

WIDDOWS Dominic 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *Proceedings of HLT/NAACL*. 276-283.

WIEMER-HASTINGS Peter & Iraide ZIPITRIA 2001. Rules for syntax, vectors for semantics. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. 1112-1117.

WITTGENSTEIN Ludwig 1953. *Philosophical Investigations*. (Translated by G.E.M. Anscombe). Oxford: Blackwell.

YAROWSKY David 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the 14th Iinternational Conference on Computational Linguistics*. Association for Computational Linguistics. 454-460.