Lexicalised compositionality

Ann Copestake
University of Cambridge

Aurelie Herbelot
Universität Potsdam

Abstract In this paper, we discuss how distributional semantics can be formally related to a simple model-theoretic approach, with a view to proposing a single account for both the phenomena traditionally covered by formal semantics and those dealt with by lexical semantics. We define some of the traditional lexical semantic relationships within this framework, and also outline its treatment of phenomena which have been considered within Generative Lexicon theory. We further discuss how the approach acknowledges linguistic differences amongst speakers of a same language.

Keywords: lexical semantics, formal semantics, distributional semantics, generative lexicon

1 Introduction

Distributional approaches to semantics are based on the principle that the linguistic contexts in which a word (or a phrase) appears can be used as a partial representation of its meaning.[1] The usual approach to formalizing and implementing this idea is to represent a word using a vector whose elements correspond to features derived from the contexts in which that word occurs. The approach can be thought of as defining a high-dimensional meaning space, and hence the terms **vector space model** or **semantic space model** are also used. In the simplest case, the features correspond to the words found within a window of text on either side of the term of interest. Various operations can be performed on these representations: for instance, by computing similarities between vectors, we can obtain a measure of the similarity between terms. While such models are now very often used in computational linguistics and psychology, and have had considerable empirical success in modelling some aspects of language, there has been relatively little discussion of them within the recent linguistics literature. One reason for this may be that distributional models are seen as 'engineering', without theoretical validity, or perhaps that they are simply too different from the approaches to semantics which have been studied within linguistics.

Our goal in this paper is to explore a theoretical approach by which we can relate distributional semantics to compositional semantics and to lexical semantics, in an attempt to build bridges rather than to replace existing accounts. We have attempted to make the account accessible to readers who may be unfamiliar with the computational literature, though we do not provide a comprehensive overview of the techniques. The core idea is to work with distributions in logical forms. For instance, where conventional logical representations might contain cat', the set of all cats in some world, we will instead use cat°, the set of all contexts in which "cat" has been uttered. As we will discuss, an equivalence can be obtained between a model-theoretic account and a distributional account. A relationship can also be seen with lexical semantics. To make these links requires a novel idealization of the notion of distribution, which we develop in this paper.

---

[1] Harris (1954) is usually cited as being the first linguist to express this idea: we discuss the history briefly in §7.

The reason for attempting to provide these bridges is that we believe that distributional approaches complement compositional semantics and non-distributional approaches to lexical semantics. We assume that a full account of semantics should support compositionality and inference, as is generally accepted in formal semantics. It should also provide a way of representing lexical meaning, including a non-stipulative approach to word senses and regular polysemy. We further assume that an account of semantics should be plausible with respect to learnability, and allow for differences between individuals in their beliefs about lexical meaning. These latter issues have not traditionally been given priority in formal accounts: in fact, we believe that the traditional Fregean view of sense leads to a dead end in these respects.

This paper belongs to a tradition in computational linguistics which takes syntax and formal semantics seriously, but which attempts to arrive at a notion of semantics which is potentially compatible with complete coverage of a language (as used in general text corpora, for instance) and which makes realistic assumptions about ambiguity. Elsewhere (Copestake 2009), one of us used the tongue-in-cheek term ‘slacker semantics’ for this approach, though many of the ideas we draw on date back at least to Hobbs (1985). What we want to argue here is that we can build on the computational semanticists’ practically-oriented approach to provide a mechanism for integrating lexical and formal semantics. This will involve an alternative underpinning to formal semantics, but one that enables us to keep intact most of the ideas that formal semantics has developed.

The hypothesis to be investigated here is that instead of talking about the set of all things in the world denoted by, say, cat′, as in an extensional account, or using a Fregean notion of sense, we talk about the **context set** for *cat*. In §2, we will introduce the idea of an **ideal distribution**, where we consider all the contexts in which *cat* could occur. Each context corresponds to the logical form of a sentence/utterance. For example, contexts where *cat* is the subject of *sleep* will be a subset of all the contexts where *cat* occurs in subject position, which in turn will be a subset of all the contexts in which *cat* occurs. We will show how this setup allows us to a) draw a direct correspondence between distributions and the standard idea of denotation and b) re-define classical lexical semantics notions and relations such as ‘sense’ or ‘hyponymy’ in terms of contexts.

Additionally, distributional context sets in our approach are specific to individual speakers. This allows different individuals to have somewhat different models of lexical concepts. Something may be a mug to one speaker and a cup to another. But speakers are also aware when concepts are borderline and are generally able to accommodate different uses, especially in grounded contexts. Someone may think of a particular object as clearly a cup, but if they are asked to ‘Pass the mug’ and that object is the only ceramic drinking vessel visible, they will generally pass it without quibbling. To allow for accommodation effects, we need to be able to compute similarity between lexemes, and distributions support this.

We should clarify here that despite the unificatory advantages of such an approach, one aspect of meaning remains unrepresented. Indeed, many utterances are directly grounded in that they refer to a situation which is evident to the hearer. This would be true of much child-directed speech, for instance. Thus we assume that some elements in the context set are paired with salient perceptual data, and that at least some of the distributional predicates can be put into correspondence with real world entities by the hearer. What we want to achieve via distributional semantics is an account of how utterances can be understood which are either not immediately grounded at all or only partially grounded. We would argue that this constitutes the vast majority of the utterances perceived by an adult. Thus the role of distributional semantics is partially to relate ungrounded words to grounded ones. For example, a hearer who has no prior knowledge of aardvarks should be able to relate aardvark$^\circ$ to known concepts without ever seeing an aardvark. We cannot (currently) simulate grounding experimentally, but if we assume some concepts are grounded, we can investigate whether our distributional techniques could result in a new ungrounded word being suitably categorised. Operations such as categorisation, similarity and paraphrase are possible (to some extent) with systems that capture relationships between words but do not emulate anything approaching real understanding, which we accept requires grounding.

In this paper, we will lay some groundwork for the idea of a context set and what it might correspond to. In the next section, we introduce the notion of the ideal distribution, which allows us to link distributional accounts directly with model theoretic accounts. We also introduce a couple of operations which, applied to context sets, will let us formally define a number of lexical relations (§3). In §4, we outline how various phenomena of lexical semantics might be analysed in our approach. In section §5, we turn to empirically observed distributions and discuss how they can be utilised. We also explain why a new type of corpus would eventually be required to build the types of models we are interested in. This discussion lets us then relate our approach to a well-known account of lexical semantics, the Generative Lexicon (Pustejovsky 1995, §6). Finally, in §7, we provide a brief survey of some of the current computational work on distributional techniques and related topics.

2 Ideal distributions

In order to make an explicit comparison with model-theoretic semantics, we will consider the hypothetical case of complete distributional information with respect to some microworld. We refer to this as an **ideal distribution**, and the particular class of ideal distributions discussed in this section as lc_0 distributions. These will be defined so that we can obtain a simple correspondence with a (first-order) notion of extension.

2.1 Ideal distributions and context sets

We will consider very simple examples with situations where the available lexemes are the adjectives *white, black*, the nouns *sphere, cube, object*, the verbs *jiggle, rotate* and the determiner *a*. We will initially consider the situation S_1 where there is a jiggling black sphere and a rotating white cube. We will call the sphere s, the jiggling event e_s, the cube c and the rotating event e_c.[2]

First we can consider the traditional approach where the denotation of predicates corresponding to the lexemes is defined in terms of sets of entities and tuples. The predicates and their denotation in S_1 are:

$$\begin{array}{l}
\text{black}' = \{s\} \\
\text{white}' = \{c\} \\
\text{sphere}' = \{s\} \\
\text{object}' = \{s, c\} \\
\text{cube}' = \{c\} \\
\text{jiggle}' = \{\langle e_s, s\rangle\} \\
\text{rotate}' = \{\langle e_c, c\rangle\}
\end{array}$$

We have the usual notion of truth, so $\text{black}'(s)$ is true and $\text{black}'(c)$ is false, for instance.

---

[2] Note that we are assuming a neo-Davidsonian account, whereby all verbal predicates have events as the first argument.

---

a sphere jiggles
a black sphere jiggles
a cube rotates
a white cube rotates
an object jiggles
a black object jiggles
an object rotates
a white object rotates

Figure 1 Sentences associated with situation S_1

For lc_0 distributions we take all possible truthful assertions using only the limited vocabulary, excluding cases where there is logical redundancy within the sentence.[3] The possible utterances corresponding to S_1 using the specified lexemes are shown in Figure 1. The "logical redundancy" condition is intended to exclude examples such as *a white white cube rotates*.

In Figure 2, we show the context sets paired with the situation described (i.e., all the utterances are grounded by S_1).

We will first discuss the form of the context sets shown in Figure 2. In our approach, the context sets for a lexeme are described in terms of logical forms (LF), one per sentence in which the lexeme occurs. We will assume relatively shallow LFs here, of the type that can be extracted reasonably efficiently and accurately from an automatic parser. In fact, we will base our analyses on those produced by the English Resource Grammar (ERG: Flickinger 2000), but simplify them for expository purposes. We distinguish between the predicate symbols corresponding to a word in the LF only if they correspond to entries which can be distinguished on syntactic grounds. For instance, we assume a single predicate including both the financial and geographic nominal senses of *bank*. Our lexemes may thus correspond to multiple word senses, even multiple homonyms. We are working with a version of Minimal Recursion Semantics (MRS: Copestake, Flickinger, Sag & Pollard 2005) representation under the general 'slacker semantics' assumption that the representation captures the information available from syntax but does not make distinctions that syntax cannot resolve.[4] MRS representations may be underspecified for certain ambiguities which are not resolved by syntax, such as scope ambiguity. An MRS structure consists of implicitly conjoined **elementary predications** consisting of a predicate and its arguments (e.g., $\text{rotate}'(e,x)$). In this section, for simplicity, we assume a 'quantifier-free' fragment of MRS, where the arguments to predicates are to be taken as constants. For instance, the sentence *a white cube rotates* results in the LF:

$$a(x4), \text{white}^\circ(x4), \text{cube}^\circ(x4), \text{rotate}^\circ(e4, x4)$$

Note that we use different argument identifiers in each LF (i.e., for each sentence): we will refer to the objects and events thus referred to as **linguistic entities**. We will discuss the grounding of the linguistic entities with respect to the actual entities in the situation below. Unlike normal

---

[3] Concentration on assertions here is motivated by the aim of showing a correspondence with the standard notion of extension. However, we believe the exclusive use of assertions is generally valid for discussion of distributional techniques, since very few words have substantially different behaviour in other speech act contexts.

[4] For computational purposes, it is also relevant that there is a variant of MRS, Dependency MRS (DMRS), which can be represented as a graph. However, we will not discuss this further here.

$$\begin{array}{lll}
\text{sphere}^{\circ} \equiv & \{ & <[x1],[a(x1),\text{jiggle}^{\circ}(e1,x1)],S_1>, \\
 & & <[x2],[a(x2),\text{black}^{\circ}(x2),\text{jiggle}^{\circ}(e2,x2)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{cube}^{\circ} \equiv & \{ & <[x3],[a(x3),\text{rotate}^{\circ}(e3,x3)],S_1>, \\
 & & <[x4],[a(x4),\text{white}^{\circ}(x4),\text{rotate}^{\circ}(e4,x4)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{object}^{\circ} \equiv & \{ & <[x5],[a(x5),\text{jiggle}^{\circ}(e5,x5)],S_1>, \\
 & & <[x6],[a(x6),\text{black}^{\circ}(x6),\text{jiggle}^{\circ}(e6,x6)],S_1>, \\
 & & <[x7],[a(x7),\text{rotate}^{\circ}(e7,x7)],S_1>, \\
 & & <[x8],[a(x8),\text{white}^{\circ}(x8),\text{rotate}^{\circ}(e8,x8)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{jiggle}^{\circ} \equiv & \{ & <[e1,x1],[a(x1),\text{sphere}^{\circ}(x1)],S_1>, \\
 & & <[e2,x2],[a(x2),\text{black}^{\circ}(x2),\text{sphere}^{\circ}(x2)],S_1>, \\
 & & <[e5,x5],[a(x5),\text{object}^{\circ}(x5)],S_1>, \\
 & & <[e6,x6],[a(x6),\text{black}^{\circ}(x6),\text{object}^{\circ}(x6)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{rotate}^{\circ} \equiv & \{ & <[e3,x3],[a(x3),\text{cube}^{\circ}(x3)],S_1>, \\
 & & <[e4,x4],[a(x4),\text{white}^{\circ}(x4),\text{cube}^{\circ}(x4)],S_1>, \\
 & & <[e7,x7],[a(x7),\text{object}^{\circ}(x7)],S_1>, \\
 & & <[e8,x8],[a(x8),\text{white}^{\circ}(x8),\text{object}^{\circ}(x8)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{black}^{\circ} \equiv & \{ & <[x2],[a(x2),\text{sphere}^{\circ}(x2),\text{jiggle}^{\circ}(e2,x2)],S_1>, \\
 & & <[x5],[a(x5),\text{object}^{\circ}(x5),\text{jiggle}^{\circ}(e5,x5)],S_1>\}
\end{array}$$

$$\begin{array}{lll}
\text{white}^{\circ} \equiv & \{ & <[x4],[a(x4),\text{cube}^{\circ}(x4),\text{rotate}^{\circ}(e4,x4)],S_1>, \\
 & & <[x8],[a(x8),\text{object}^{\circ}(x8),\text{rotate}^{\circ}(e8,x8)],S_1>\}
\end{array}$$

Figure 2 Ideal context sets for S_1

MRS, we notate the predicates corresponding to the open-class lexemes in this sentence using the notation P°. In general, we will assume a distributional interpretation for open class words and a non-distributional meaning for closed class words (*a* in this example).[5]

We define the context set for a lexeme *l* in terms of the logical forms which contain an elementary predication corresponding to *l*. [6] We will refer to the set of such logical forms as LF(l). An element in the context set for *l* derived from a logical form *lf* which is a member of LF(l) consists of a pair of a **distributional argument tuple** and a **distributional LF** $< args, dlf >$ where the distributional arguments *args* are the arguments associated with the elementary predication corresponding to *l* in *lf*, and *dlf* is *lf* with that elementary predication removed. In the case of the sentence *a white cube rotates*, this gives the context set element

$$< [x4], [a(x4), \mathrm{cube}^\circ(x4), \mathrm{rotate}^\circ(e4, x4)] >$$

in the distribution white°. For the grounded utterances, we pair the context set elements with the corresponding situations, giving:

$$< [x4], [a(x4), \mathrm{cube}^\circ(x4), \mathrm{rotate}^\circ(e4, x4)], S_1 >$$

The full context set contains all the elements corresponding to the lexeme *l*. As should be evident from Figure 2, a single sentence will generally correspond to multiple context set elements, one for each open class lexeme which it contains.

2.2 Context sets and extensions

There is a very straightforward correspondence between the lc_0 context sets and the standard notion of extension under the assumption that the equalities between the constants corresponding to distributional arguments are known. For instance, consider the distributional arguments for sphere° and object° and assume that we know $x1 =_{rw} x2 =_{rw} x5 =_{rw} x6 =_{rw} s$ and that $x7 =_{rw} x8 =_{rw} c$ (where $=_{rw}$ stands for real world equality):

$$\begin{array}{lll}
\mathrm{sphere}^\circ \equiv & \{ & < [s], [a(s), \mathrm{jiggle}^\circ(e_s, s)], S_1 >, \\
 & & < [s], [a(s), \mathrm{black}^\circ(s), \mathrm{jiggle}^\circ(e_s, s)], S_1 > \} \\
\mathrm{object}^\circ \equiv & \{ & < [s], [a(s), \mathrm{jiggle}^\circ(e_s, s)], S_1 >, \\
 & & < [s], [a(s), \mathrm{black}^\circ(s), \mathrm{jiggle}^\circ(e_s, s)], S_1 >, \\
 & & < [c], [a(c), \mathrm{rotate}^\circ(e_c, c)], S_1 >, \\
 & & < [c], [a(c), \mathrm{white}^\circ(c), \mathrm{rotate}^\circ(e_c, c)], S_1 > \}
\end{array}$$

Thus the distributional arguments of P° in lc_0 correspond to P'.

The condition for this correspondence is that for each situation entity z, for every predicate P' for which $\mathrm{P}'(z)$ is true, we have a logical form for a sentence in the lc_0 distribution containing an elementary predication equivalent to $\mathrm{P}^\circ(z)$. We do not actually need all the sentences shown in Figure 1 to establish the equivalence. However, we want to use the idea of "all sentences

---

[5] This is an approximation: for instance, there are arguments for treating prepositions distributionally in some contexts. However, we will not explore this further here.

[6] For simplicity here, we will only consider the cases where there is just one such elementary predication in the LF.

---

a cube rotates
a black cube rotates
an object rotates
a black object rotates

Figure 3 Sentences corresponding to the situation S_2

---

corresponding to a situation S" rather than talk about truth conditions as in a conventional model-theoretic approach because we want the lc_0 concept to be intuitively meaningful by itself and not to rely on the standard notion of denotation.

Linking the linguistic entities to the entities in the situation requires some knowledge of the relationship between the utterances and situations but does not require that the hearer has full knowledge of lexical meaning. Assume that a language learner perceives S_1 and the associated sentences, is capable of producing the LFs but is not aware of the meaning of the open class lexemes. We also assume that the learner can distinguish objects from events and has an expectation that different nouns refer to different entities unless they have evidence to the contrary, which is consistent with the psycholinguistic evidence on language learning (see, e.g., Carey 2009). Under these assumptions, given the context sets in Figure 2, the learner will always assign $x1 =_{rw} x2 =_{rw} x5 =_{rw} x6$, $e1 =_{rw} e2 =_{rw} e5 =_{rw} e6$, $x3 =_{rw} x4 =_{rw} x7 =_{rw} x8$ and $e3 =_{rw} e4 =_{rw} e7 =_{rw} e8$ but might assign the groups to the wrong entities and events.
Correct assignment:

$$\begin{aligned} &x1 =_{rw} x2 =_{rw} x5 =_{rw} x6 =_{rw} s \\ &e1 =_{rw} e2 =_{rw} e5 =_{rw} e6 =_{rw} e_s \\ &x3 =_{rw} x4 =_{rw} x7 =_{rw} x8 =_{rw} c \\ &e3 =_{rw} e4 =_{rw} e7 =_{rw} e8 =_{rw} e_c \end{aligned}$$

Incorrect assignment:

$$\begin{aligned} &x1 =_{rw} x2 =_{rw} x5 =_{rw} x6 =_{rw} c \\ &e1 =_{rw} e2 =_{rw} e5 =_{rw} e6 =_{rw} e_c \\ &x3 =_{rw} x4 =_{rw} x7 =_{rw} x8 =_{rw} s \\ &e3 =_{rw} e4 =_{rw} e7 =_{rw} e8 =_{rw} e_s \end{aligned}$$

However, the correct assignment can be identified if further information is available. Consider an additional situation S_2 where there is a black cube ($c1$) which is rotating (e_{c1}). The sentences corresponding to S_2 are shown in Figure 3. Figure 4 shows the combined lc_0 distributions for the two situations. Given that there is only one entity and one event in S_2, the identities $x9 =_{rw} x10 =_{rw} x11 =_{rw} x12 =_{rw} c1$ and $e9 =_{rw} e10 =_{rw} e11 =_{rw} e12 =_{rw} e_{c1}$ are trivially established. Now assuming only that the e_{c1} event is perceptually more similar to e_c than to e_s, the learner can identify the correct assignment in S_1. The distributions and the identification of the linguistic entities with the 'real world' entities can thus proceed via comparison without any sort of explicit meaning being associated with the lexemes. These properties are attractive for an account of semantics which supports a realistic model of language learning.

It is straightforward to derive distributions for phrases, such as black_sphere$^\circ$ by treating them in the same way as lexemes. It should be clear that the distribution for *black sphere* can also be related to the intersection of the context sets for black$^\circ$ and sphere$^\circ$. Note that this does not rely on

$$\begin{array}{lll}
\text{sphere}^\circ \equiv & \{ & <[x1],[a(x1),\text{jiggle}^\circ(e1,x1)],S_1>, \\
& & <[x2],[a(x2),\text{black}^\circ(x2),\text{jiggle}^\circ(e2,x2)],S_1>\} \\
\text{cube}^\circ \equiv & \{ & <[x3],[a(x3),\text{rotate}^\circ(e3,x3)],S_1>, \\
& & <[x4],[a(x4),\text{white}^\circ(x4),\text{rotate}^\circ(e4,x4)],S_1>, \\
& & <[x9],[a(x9),\text{rotate}^\circ(e9,x9)],S_2>, \\
& & <[x10],[a(x10),\text{black}^\circ(x10),\text{rotate}^\circ(e10,x10)],S_2>\} \\
\text{object}^\circ \equiv & \{ & <[x5],[a(x5),\text{jiggle}^\circ(e5,x5)],S_1>, \\
& & <[x6],[a(x6),\text{black}^\circ(x6),\text{jiggle}^\circ(e6,x6)],S_1>, \\
& & <[x7],[a(x7),\text{rotate}^\circ(e7,x7)],S_1>, \\
& & <[x8],[a(x8),\text{white}^\circ(x8),\text{rotate}^\circ(e8,x8)],S_1>, \\
& & <[x11],[a(x11),\text{rotate}^\circ(e11,x11)],S_2>, \\
& & <[x12],[a(x12),\text{black}^\circ(x12),\text{rotate}^\circ(e12,x12)],S_2>\} \\
\text{jiggle}^\circ \equiv & \{ & <[e1,x1],[a(x1),\text{sphere}^\circ(x1)],S_1>, \\
& & <[e2,x2],[a(x2),\text{black}^\circ(x2),\text{sphere}^\circ(x2)],S_1>, \\
& & <[e5,x5],[a(x5),\text{object}^\circ(x5)],S_1>, \\
& & <[e6,x6],[a(x6),\text{black}^\circ(x6),\text{object}^\circ(x6)],S_1>\} \\
\text{rotate}^\circ \equiv & \{ & <[e3,x3],[a(x3),\text{cube}^\circ(x3)],S_1>, \\
& & <[e4,x4],[a(x4),\text{white}^\circ(x4),\text{cube}^\circ(x4)],S_1>, \\
& & <[e7,x7],[a(x7),\text{object}^\circ(x7)],S_1>, \\
& & <[e8,x8],[a(x8),\text{white}^\circ(x8),\text{object}^\circ(x8)],S_1>, \\
& & <[e9,x9],[a(x9),\text{cube}^\circ(x9)],S_2>, \\
& & <[e10,x10],[a(x10),\text{black}^\circ(x10),\text{cube}^\circ(x10)],S_2>, \\
& & <[e11,x11],[a(x11),\text{object}^\circ(x11)],S_2>, \\
& & <[e12,x12],[a(x12),\text{black}^\circ(x12),\text{object}^\circ(x12)],S_2>\} \\
\text{black}^\circ \equiv & \{ & <[x2],[a(x2),\text{sphere}^\circ(x2),\text{jiggle}^\circ(e2,x2)],S_1>, \\
& & <[x6],[a(x6),\text{object}^\circ(x6),\text{jiggle}^\circ(e6,x6)],S_1>, \\
& & <[x10],[a(x10),\text{cube}^\circ(x10),\text{rotate}^\circ(e10,x10)],S_2>, \\
& & <[x12],[a(x12),\text{object}^\circ(x12),\text{rotate}^\circ(e12,x12)],S_2>\} \\
\text{white}^\circ \equiv & \{ & <[x4],[a(x4),\text{cube}^\circ(x4),\text{rotate}^\circ(e4,x4)],S_1>, \\
& & <[x7],[a(x7),\text{object}^\circ(x7),\text{rotate}^\circ(e7,x7)],S_1>\}
\end{array}$$

Figure 4 Ideal context sets for Situations 1 and 2 grounding the linguistic entities. While there is much recent work in computational linguistics on appropriate vector space models for phrases, which we briefly discuss in §7, we do not need these for our theoretical account of meaning for compositional phrases.[7]

Because the conventional concept of logical denotation can be derived from the lc_0 distributions, we can define a standard notion of logical inference. Quantifiers can be defined in terms of the real world entities. But we can also see how some inferences are possible on the basis of the distributions alone. Hyponymy relationships correspond to a subset relationship between context sets modulo argument renaming: e.g., cube$^\circ$ is a subset of object$^\circ$ in Figure 4. Synonyms would have equal context sets (again, modulo argument names). Note that, in order to get such inclusion relationships with full quantified LFs, we must process quantified statements before adding them to the ideal distribution. We discuss these matters in more detail in Sections 3.2 (on quantifiers) and 4 (on lexical relations).

2.3 Linguistic entities and reference

Before going into details of how our notion of context set is related to more usual accounts of distribution, we will elaborate a little on our notion of a linguistic entity, in which we are essentially following the approach advocated by Hobbs (1985). The level of indirection provided by distinguishing between linguistic entities and real world entities has a number of advantages from our viewpoint, as illustrated by the example we gave of a learner distinguishing between *sphere* and *cube*. In fact, although we sometimes loosely use the term ‘real world entities’ instead of ‘referent’, we are not interested in whether the situation grounding an utterance corresponds to the real world or a fictional one. There is no issue of whether something actually exists in the real world or not at the distributional level: unicorns have the same status as cats.

Our notion of intension corresponds to the context sets of lexemes in the ideal distributions. This implies that there will be multiple linguistic concepts which are real world identical. This allows us to dodge (or postpone) many standard puzzles. The Morning Star and Evening Star will be different linguistic concepts, and a speaker may or may not be aware that these map to the same real world entity. Mappings to real world concepts may change without affecting the linguistic concepts substantially: for instance, the distribution of *tiger* will not substantially change if it suddenly turns out they are all Martian robots. If Kim, who is both judge and hangman, is on strike as a judge, we would not necessarily expect *the hangman is on strike* to occur in the ideal distribution. Finally, speakers do not necessarily appreciate logical consequences of mappings to the real world. This general approach naturally gives rise to a different set of difficulties, in particular how an individual develops and updates concepts, but the attraction is that these problems relate much more clearly to research on psychology (e.g., Carey 2009). In fact, this line may be of interest even in highly formal uses of language: Ganesalingam (2009) suggests that modelling concept change may be crucial to analysing the language of mathematics. Of course, making this argument properly would require a detailed discussion: the point we want to make here is just that we believe that distinguishing between linguistic entities and referents is more than just a convenient computational linguistics hack.

---

[7] Multiword expressions (MWEs) require a different approach. Our notion of LF for the context sets is based on the assumption that non-compositional multiword expressions have their own lexical entries and can be treated as giving rise to a single predicate symbol. For instance, a verb-particle such as *run up* in *Kim ran a large bill up* would correspond to run_up$^\circ$.

	sphere	jiggles	black	cube	rotates	white	object
sphere	–	1	1	0	0	0	0
jiggles	1	–	1	0	0	0	1
black	1	1	–	0	0	0	1
cube	0	0	0	–	1	1	0
rotates	0	0	0	1	–	1	1
white	0	0	0	1	1	–	1
object	0	1	1	0	1	1	–

Figure 5 Binary distributional vectors derived from sentences in Figure 1.

	sphere	jiggles	black	cube	rotates	white	object
sphere	–	2	1	0	0	0	0
jiggles	2	–	2	0	0	0	2
black	1	2	–	0	0	0	1
cube	0	0	0	–	2	1	0
rotates	0	0	0	2	–	2	2
white	0	0	0	1	2	–	1
object	0	2	1	0	2	1	–

Figure 6 Basic distributional vectors with counts derived from sentences in Figure 1.

2.4 Contexts and vectors

We now turn to discussing how the context can be treated in terms of vectors, as in more standard approaches to distributional semantics. The most basic approach to distributional semantics uses a vector representation of the context expressed in terms of individual words (or lexemes). For instance, assuming that the context is the individual sentence in which a word appears, the sentences shown in Figure 1 would give the binary vector shown in Figure 5 (the vector elements record the presence or absence of a word in the context) or the integer vector in Figure 6 (elements record the counts). We have omitted *a/an*, as it is usual to exclude some very common words from the distributions. There is a large range of possible representations described in the computational literature, which we will not attempt to summarise here.

In our approach, the elements of the vector are components of the context sets, but there are a number of options as to exactly what the components are. If we take all the individual predications in the context set (the elementary predications in MRS terms), the components include predications which are not directly related to the term under consideration, as in the simplest approaches to distributional semantics. For example, the distribution for ‘jiggle’ based on the context set corresponding to ‘the ball on the table jiggled’ would include table$'(x)$. On the other hand, we might only be interested in predications which directly relate to an entity corresponding to the word under consideration. In this case, table$'(x)$ would be omitted, since it would not be directly related to a jiggling event. Of course, we could decide to include predications which are related by paths of predications of up to a certain length, or only include paths of a particular type (cf Padó & Lapata

	$a(x)$	black$^\circ(x)$	white$^\circ(x)$	jiggle$^\circ(e,x)$	rotate$^\circ(e,x)$	sphere$^\circ(x)$	cube$^\circ(x)$	object$^\circ(x)$
sphere$^\circ$	1	1	0	1	0	0	0	0
cube$^\circ$	1	0	1	0	1	0	0	0
object$^\circ$	1	1	1	1	1	0	0	0
black$^\circ$	1	0	0	1	0	1	0	1
white$^\circ$	1	0	0	0	1	0	1	1

Figure 7 Vectors corresponding to context sets for S_1

2007). We also have a choice as to what level of decomposition we apply, since we could make the elements of the vector correspond to single predications only (e.g., black$^\circ(x)$) or also include groupings of predications (e.g., black$^\circ(x)$,jiggle$^\circ(e,x)$).

Vectors corresponding to the ideal context sets for S_1 are shown in Figure 7 (which should be compared to Figure 2). For this example, we have assumed single predications which directly relate to the lexeme being considered. To make the figure more readable, we have omitted the context sets for the verbs and predications relating to events (e.g., $[e]$jiggle$^\circ(e,x)$) and assumed all predications relate to x (in the full representation, this has to be explicit and there will be two components corresponding to 'jiggle' for instance: $[e]$jiggle$^\circ(e,x)$ and $[x]$jiggle$^\circ(e,x)$). The components in the vector correspond to simple predications. The 'flat' MRS representation means that the decomposition of the semantic representation into elementary predications is trivial. We are glossing over the precise formulation of the transformation of the context sets into vectors here, but will return to this issue in §3.

The vector representation is a way of generalising over the elements in the context sets. If directly-connected predications are assumed, then the elements can be thought of as corresponding to a very fine-grained notion of semantic feature. The more general words, such as *object*, provide a way of generalising over the more specific features. In this very contrived setting, for instance, black$^\circ$ and white$^\circ$ only share the *a* and *object* contexts. If we had included *move* in the vocabulary as a generalisation of *jiggle* and *rotate*, the vector would provide a means of separating movable and immovable entities. Further generalisations would be possible with the use of a more decomposed logical form, with an explicit representation of roles. For example, we could have jiggle$^\circ(e)$,$ARG1(e,x)$ instead of jiggle$^\circ(e,x)$: this style of representation would allow a separation to be made between entities which occurred in subject position and those that did not.

2.5 Context set subspaces

We establish context sets at the level of lexemes, with each lexeme being represented by a full context set, as illustrated in Figures 2 and 4. We can also consider various subspaces of the context set by considering different parts of the vectors. In theory, any subspace can be distinguished in a distribution but most have no linguistic relevance and are therefore of no interest to us. However, some subspaces relate to standard linguistic concepts. In particular, the conventional notion of a word sense should correspond to a relatively homogeneous subspace of a lexeme's context set, although we would argue that it is generally impossible to precisely delimit such subspaces. For instance, the distributional subspaces that are part of bank$^\circ$ would be distinguished because the other predicates contained in the distributional LF differ. The financial *bank* might be associated with *lend*, *overcharge* and *bankrupt* while the geographical feature is associated with *sandy*, *picnic* and *otter*. A range of approaches to deriving sense clusters from distributions have been described in the computational linguistics literature (see, e.g., Schütze 1998, Lin & Pantel 2002, Thater, Fürstenau & Pinkal 2011). In general, clear cases of homonymy, such as the bank example, give rise to relatively discrete clusters.

We note here that in our approach these subspaces will be associated with sets of linguistic entities with negligible overlap. Although there are some predicates which are associated with both senses of *bank* (e.g., *collapse*), we would not expect to find utterances where e.g., *sandy* and *overcharge* are applied to the same linguistic entity. In section §6, we will contrast this with examples such as *book*, where predicates that relate to intuitively different subspaces can both be used of the same entity.

Individual entities will correspond to finer-grained subspaces. In Figure 4, cube$^\circ$ contains subspaces corresponding to two different referents: one denoted by the constants $x3$ and $x4$, which correspond to the entity we called c, and one denoted by $x9$ and $x10$, which we called $c1$. So, for instance, the distribution of *cat* in the sense of a small furry animal contains many subspaces which correspond to various individual cats, each one with its own distribution; selecting one entity out of the cat-meaning-animal subspace means selecting one of those distributions.

In the trivial examples shown, we have only discussed singular terms. We can extend these ideas to plurals by assuming that a plurality is a sum of individuals, as described by Link (1983). We assume a Linkian view of plurals as join-semi-lattices where each point at the bottom of the lattice corresponds to one entity and all other points are sums of singular entities, or sums of sums.[8] So a plurality corresponds to a subspace which comprises two or more entities which are themselves subspaces of that plurality. Note that in general, we cannot say that the distribution of a plurality is the union of the distributions of its individual entities. A plural distribution will also include contexts that apply only to the sum of individuals and not to the individuals themselves (i.e., collective, as opposed to distributive, contexts).

In this section, we have argued that distributions could potentially form the basis of a general approach to word meaning. Of course, this notion of an ideal distribution is a largely hypothetical exercise. We will not, for instance, see a subset relationship between cube$^\circ$ and object$^\circ$ in real data. However, we think that ideal distributions have a psychological reality in that they refer to the 'linguistic potential' of an individual, that is, the utterances that they might produce in response to a stimulus given their knowledge of the situation (they may not know that the rotating cube is hiding a motionless sphere), the vocabulary available to them and their linguistic beliefs (e.g., whether they describe objects of a particular shape as *mug* or *cup*). We also think that the notion can act as a guide in considering how we model the relationship between what an individual is exposed to (the **actual distributions**) and the individual's internal **language model**. We explore this in more detail in Section 5.

3 Operations on ideal distributions

Having introduced the notion of ideal distribution, we will consider some phenomena of traditional lexical semantics and their translation in LC. Building formal definitions for those phenomena,

---

[8] We will not discuss mass terms here, but in principle, we accept Chierchia's revision of Link's view (Chierchia 1998), where mass terms consist of minimal parts.

however, implies the ability to perform certain operations over ideal context sets. In what follows, we will introduce two such operations: a) how to reduce ideal distibutions to an underspecified form and b) how to deal with quantified logical forms.

3.1 Underspecified context sets

Context sets as described in §2 include information about instances and situations. By reducing a context set to a representation that includes logical forms only, we get a derived distributional form which is more akin to the linguistic objects typically assumed by computational linguists, but which preserves some of the of the properties of the ideal distribution – crucially, what we will call its **information saturation** property (i.e. the fact that there is a logical form for each event/relation in the world under consideration).

Consider the following three contexts in the distribution of *cat* and assume $x1 =_{rw} x11$ but $x1 \neq_{rw} x2$:

$$< [x1], [a(x1), \text{sleep}^\circ(e1,x1)], S_1 >$$
$$< [x11], [a(x11), \text{sleep}^\circ(e2,x11)], S_2 >$$
$$< [x2], [a(x2), \text{sleep}^\circ(e3,x2)], S_3 >$$

It is possible to underspecify those contexts with respect to entities and situations by writing:

$$< [x], [a(x), \text{sleep}^\circ(e,x)], S >$$
$$< [x], [a(x), \text{sleep}^\circ(e,x)], S >$$
$$< [x], [a(x), \text{sleep}^\circ(e,x)], S >$$

The above contexts tell us that we have three distinct entity tuples involving some cat instance with some sleeping event in some situation. It is unknown whether it is the same cat involved in all events, or three cats, or two, and how many sleeping events and situations are implied.

Generally, we can define the underspecified form of an ideal context set 1° as a set $\mathscr{U}(l)$ of tuples $(lf, arg_{1...n}, S)$, where lf is a logical form, S is a situation and $arg_{1...n}$ are lf's arguments. Knowing the correspondence between $e_{1...n}$ and the distributional arguments in 1°, and between S and the situations in 1°, allows to return to the fully specified form of the ideal context set.

We will refer to the underspecified form of 1° as $1^\circ_{\mathscr{U}}$.

3.2 Quantification: unpacking distributions

Our notion of ideal distribution presupposes a direct correspondence to set-theoretical models where each distributional argument for a logical form corresponds to one, *and only one*, individual in the world under consideration, i.e. to a point in a set, and is accordingly singularly quantified in the logical form. This implies that plurally quantified statements must be appropriately converted before being included in an ideal context set.

We define the process of **unpacking** as the translation of a logical form containing non-individually quantified arguments into several logical forms, one for each element in the set denoted by the quantified argument:

For instance, in cat$^\circ$ we might have:

$$< [x1][three(x1), \mathrm{sleep}^\circ(e1,x1)], S_1 >= \{ \begin{array}{l} < [x11][one(x11), \mathrm{sleep}^\circ(e11,x11)], S_1 >, \\ < [x12][one(x12), \mathrm{sleep}^\circ(e12,x12)], S_1 >, \\ < [x13][one(x13), \mathrm{sleep}^\circ(e13,x13)], S_1 > \} \end{array}$$

In the unpacked representation, the quantifier becomes redundant so we can simply write:

$$< [x1][three(x1), \mathrm{sleep}^\circ(e1,x1)], S_1 >= \{ \begin{array}{l} < [x11][\mathrm{sleep}^\circ(e11,x11)], S_1 >, \\ < [x12][\mathrm{sleep}^\circ(e12,x12)], S_1 >, \\ < [x13][\mathrm{sleep}^\circ(e13,x13)], S_1 > \} \end{array}$$

We have defined the ideal distribution as a case where, with respect to a world, we have complete distributional information. In that case, it is no more difficult to unpack a universal quantifier than it is to unpack a cardinal. Unpacking simply consists of constructing the relevant distributional equality between a logical form and the set of singularly quantified logical forms containing the relevant arguments. Note that, quantifier aside, all logical forms are supposed to be identical and the plurally quantified argument denotes the same plurality as the set of all singularly quantified arguments. For instance, in a world with four cats:

$$< [x1][all(x1), \mathrm{sleep}^\circ(e1,x1)], S_1 >= \{ \begin{array}{l} < [x11][\mathrm{sleep}^\circ(e11,x11)], S_1 >, \\ < [x12][\mathrm{sleep}^\circ(e12,x12)], S_1 >, \\ < [x13][\mathrm{sleep}^\circ(e13,x13)], S_1 >, \\ < [x14][\mathrm{sleep}^\circ(e14,x14)], S_1 > \} \end{array}$$

We can proceed similarly for all quantifiers that express a ratio with respect to the universal quantifier. In the ideal distribution, we know which individuals are quantified over by *most* or *few*. For the case of collective statements, we consider the collective as a single entity.

4 Lexical semantics and ideal distributions

Traditional lexical semantics allows us to define a number of standard relations in terms of extension. For instance, full synonymy implies set identity while hyponymy can be translated into a set inclusion relation (we provide more detail on this in what follows). In contrast, it is difficult to formally (or even less formally) describe such relations in a distributional setting. Some attempts have been made to extract hyponyms from distributions, or distinguish near-synonyms from antonyms (see Section 7). Such work, however, often relies on heuristics which, although they are still in some sense distributional (i.e. they use patterns found in real text), do not define actual relations between distributions. Attempts to do so make use of machine learning techniques which, although successful (Baroni, Bernardi, Do & Shan 2012), do not result in a formal definition of the relation they are extracting but in a classifier-dependent prototypical pattern.

In this section, we will describe how standard relations in lexical semantics can be formally expressed using our notion of ideal distributions. It should be intuitively clear that, because of the direct relation between ideal distributions and extension, it is possible to retain all classical definitions of lexical relations, but we will now explicitely show how, and give formalisations in terms of LC distributions.

In what follows, we assume unpacked ideal distributions, partitioned into appropriate subspaces (see §2.5): when we talk of cube$^\circ$, we talk of the distribution of a particular subspace, or ‘sense’ in the classical account, of *cube*. The subspace we intend should be obvious from the context. We also regard fixed expressions as words with spaces, which have separate distributions from their components. We assume, for example, that for an individual who understands *to kick the bucket* as *to die*, the phrase only belongs to bucket$^\circ$ in its compositional meaning of hitting a bucket with one's foot.

4.1 Similarity

Before discussing the LC definitions of standard lexical relations, we will briefly account for the phenomenon underlying them all, i.e. similarity. Although a vague notion, similarity has been found to be a meaningful concept in psycholinguistics experiments. Miller & Charles (1991), for instance, repeating part of an experiment initially devised by Rubenstein & Goodenough (1965), showed that humans agree strongly when asked to rate the similarity of word pairs.

We define the following two notions:

- The shared distribution of two lexical items, A$^\circ$ and B$^\circ$, which may be underspecified or not:

  (1) $$S(\mathrm{A}^\circ,\mathrm{B}^\circ)=\mathrm{A}^\circ\cap\mathrm{B}^\circ$$

- The characteristic distribution of one lexical item with respect to another one:

  (2) $$C(\mathrm{A}^\circ/\mathrm{B}^\circ)=\mathrm{A}^\circ-(\mathrm{A}^\circ\cap\mathrm{B}^\circ)$$

We can give numerical values corresponding to these relations. Let us define $S_n(\mathrm{A}^\circ,\mathrm{B}^\circ)$, which expresses the degree to which A and B share contexts. Such value can be computed in a variety of ways, the simplest approach being perhaps the Jaccard metric:

(3) $$S_n(\mathrm{A}^\circ,\mathrm{B}^\circ)=\frac{|\mathrm{A}^\circ\cap\mathrm{B}^\circ|}{|\mathrm{A}^\circ\cup\mathrm{B}^\circ|}$$

Similarly,

(4) $$C_n(\mathrm{A}^\circ/\mathrm{B}^\circ)=\frac{|\mathrm{A}^\circ-(\mathrm{A}^\circ\cap\mathrm{B}^\circ)|}{|\mathrm{A}^\circ\cup\mathrm{B}^\circ|}$$

We follow Harris (1954) in his claim that lexical items that appear in the same type of contexts are semantically similar. However, due to the nature of our distributions, which include specific information about instances and situations, we must qualify this statement further.

Consider, for instance, the concepts of *cat* and *dog*. They are fairly similar, but they are never substitutable in any given existentially quantified context: we cannot point to a cat and say *This is a dog*. In fact, their shared distribution $S(\mathrm{cat}^\circ,\mathrm{dog}^\circ)$ is the empty set and $S_n(\mathrm{cat}^\circ,\mathrm{dog}^\circ)$ is 0. The overlap $S(\mathrm{cat}^\circ_{\mathscr{U}},\mathrm{dog}^\circ_{\mathscr{U}})$ of their underspecified context sets, however, can be expected to be high.

The point illustrated here is that we must differentiate between a certain notion of contextual similarity, (the one implied by Harris), which is related to selectional preference and intension, from full linguistic substitutability, which is related to real-world entities, or extension. The latter can be captured from full ideal context sets while the former must be defined in terms of underspecified context sets. The broad concept of similarity investigated by Rubenstein & Goodenough (1965) and Miller & Charles (1991) is the one also intended by Harris and we will therefore write the similarity of two lexical items A° and B° as:

(5) $$Sim(\mathrm{A}^\circ,\mathrm{B}^\circ) = S_n(\mathrm{A}^\circ_{\mathscr{U}},\mathrm{B}^\circ_{\mathscr{U}}).$$

4.2 Synonymy

Synonymy can be defined both via contextual similarity and substitutability. If two words, in a particular sense, can be substituted for each other (in both directions), in all contexts relevant to the sense under consideration, they can be called synonyms. Naturally, they are also contextually similar.

Synonymy is to some extent gradable: some words share a lot of their meaning but not all of it and are therefore not fully substitutable (for example *off* and *rancid*, where the latter is only applicable to fatty food). Sometimes words are definitionally substitutable but they present a difference in meaning which is more stylistic or emotive: see for instance *policeman/policewoman* versus *cop*. In the following, we will distinguish between true synonyms like *aubergine/eggplant*, which share their whole meanings, and near-synonyms like *rancid/off*. We will simply talk of synonyms to encompass both types.

4.2.1 True synonymy

True synonymy is a relation that must be defined using full context sets rather than underspecified context sets. In the model-theoretic framework, true synonyms are words which denote the same entities in a world (and not separate entities that happen to be extremely similar). Consequently, it is not sufficient to say that two synonyms have the same underspecified context sets: they must apply to the same situations. In our ideal setting with full distributional information, real synonymy corresponds to the complete overlap of two full context sets.

A and B are true synonyms iff

(6) $$\mathrm{A}^\circ = \mathrm{B}^\circ$$

By extension,

(7) $$S(\mathrm{A}^\circ,\mathrm{B}^\circ) = \mathrm{A}^\circ = \mathrm{B}^\circ$$

(8) $$C(\mathrm{A}^\circ/\mathrm{B}^\circ) = C(\mathrm{B}^\circ/\mathrm{A}^\circ) = \emptyset$$

(9) $$S_n(\mathrm{A}^\circ,\mathrm{B}^\circ) = 1$$

(10) $$C_n(\mathrm{A}^\circ/\mathrm{B}^\circ) = C_n(\mathrm{B}^\circ/\mathrm{A}^\circ) = 0$$

Intuitively, we can say that in a given situation s_k involving an instance a_k of A, a_k can equally be referred to using either A or B, and thus any logical form describing a_k in s_k will be contained in both the distributions of A and B.

Note that according to this definition, *policeman* and *cop* are full synonyms. The difference in their intension is not expressed in terms of ideal context sets. It could however be captured in terms of ‘actual’ distributions (see §5 for more details).

From our definition, it naturally follows that:

$$S(\mathrm{A}^{\circ}_{\mathscr{U}},\mathrm{B}^{\circ}_{\mathscr{U}}) = \mathrm{A}^{\circ}_{\mathscr{U}} = \mathrm{B}^{\circ}_{\mathscr{U}} \tag{11}$$

$$C(\mathrm{A}^{\circ}_{\mathscr{U}}/\mathrm{B}^{\circ}_{\mathscr{U}}) = C(\mathrm{B}^{\circ}_{\mathscr{U}}/\mathrm{A}^{\circ}_{\mathscr{U}}) = \emptyset \tag{12}$$

$$S_n(\mathrm{A}^{\circ}_{\mathscr{U}},\mathrm{B}^{\circ}_{\mathscr{U}}) = 1 \tag{13}$$

$$C_n(\mathrm{A}^{\circ}_{\mathscr{U}}/\mathrm{B}^{\circ}_{\mathscr{U}}) = C_n(\mathrm{B}^{\circ}_{\mathscr{U}}/\mathrm{A}^{\circ}_{\mathscr{U}}) = 0 \tag{14}$$

4.2.2 Near-synonyms

Near-synonymy is a phenomenon more related to similarity than to synonymy itself. In simple terms, it expresses ‘high similarity’. Therefore, we define it using underspecified context sets.

If *A* and *B* are near-synonyms, then

$$S_n(\mathrm{A}^{\circ}_{\mathscr{U}},\mathrm{B}^{\circ}_{\mathscr{U}}) > \delta \text{ where } \delta \text{ is ‘large’ (i.e. close to 1).} \tag{15}$$

A and *B* are near-synonyms iff Equation 15 holds *and A* and *B* are not antonyms (see antonymy definitions in §4.4).

4.3 Hyponymy

Hyponymy, or **hyperonymy**, is usually described in terms of the relationship between a more general and a more specific term: for instance, *poodle* and *dog* are two terms that can be used to describe the same entity but the former is more specific than the latter. We can also say that the extension of the more general term includes the extension of the more specific one (the set of all poodles is included in the set of all dogs). Conversely, the intension of *dog* is included in the intension of *poodle*: i.e., everything that can be said of a dog can be said of a poodle. It has been remarked, however, that the intensional definition is only applicable in an essentialist framework, where ‘dogness’ can be reduced to some essential features. What those features should be remains a puzzle: Geeraerts (2010) illustrates the issue by showing that flying cannot be an essential feature of birds if we want penguins to be birds.

We have already seen in Section 2 that in ideal distributions, an inclusion relationship can be observed between hypernyms and hyponyms. For instance, we assume cube$^{\circ}$ to be a subset of object$^{\circ}$ (see Figure 4). Intuitively, any entity can be described in terms of its hypernyms so any predicate in a logical form can be substituted for the corresponding hypernym and the full context set of the hypernym includes all logical forms found in its hyponyms. Generally, *A* is a hyponym of *B* iff:

$$\mathrm{A}^{\circ} \subset \mathrm{B}^{\circ} \tag{16}$$

It follows that:

$$\mathrm{A}^{\circ}_{\mathscr{U}} \subset \mathrm{B}^{\circ}_{\mathscr{U}} \tag{17}$$

4.4 Antonymy

Geeraerts (2010), following Lyons (1977) and Lehrer (2002), distinguishes between three basic types of antonymy: gradable, non-gradable and multiple antonyms. The gradable type refers to pairs of terms that describe opposite ends of a scale, for instance *cold* and *hot*. Such terms can be modified with adverbs of intensity such as *very* or *slightly*. Non-gradable antonyms are those that express a discrete, binary opposition like *dead* and *alive*. No scale is involved (we can't express various degrees of 'deadness', at least in the main use of *dead*) and modification is therefore unfelicitous. The last class, multiple antonyms, refers to terms that denote several discrete points on a non-gradable, discontinuous scale: traditional British academic positions (*lecturer*, *reader*, *professor*) are an example of such a scale.

Regardless of the type considered, we can define antonymy as having the following two features: firstly, it is not possible to apply antonyms to the same entity in the same situation (for instance, in *S*, it is not possible to utter *Cube X rotates clockwise* and *Cube X rotates anticlockwise*), so in terms of extension, antonyms are fully exclusive, and secondly, antonyms concern a certain concept (temperature, life and academic career in the examples above) and are therefore related in terms of intension.

In LC, if *A* and *B* are antonyms, then

$$S(\mathrm{A}^\circ,\mathrm{B}^\circ) = \emptyset \tag{18}$$

$$C(\mathrm{A}^\circ/\mathrm{B}^\circ) = \mathrm{A}^\circ \tag{19}$$

$$C(\mathrm{B}^\circ/\mathrm{A}^\circ) = \mathrm{B}^\circ \tag{20}$$

$$S_n(\mathrm{A}^\circ,\mathrm{B}^\circ) = 0 \tag{21}$$

$$C_n(\mathrm{A}^\circ/\mathrm{B}^\circ) = C_n(\mathrm{B}^\circ/\mathrm{A}^\circ) = 1 \tag{22}$$

The above formulas are, however, not sufficient to define antonymy. cat$^\circ$ and dog$^\circ$ satisfy those conditions without being antonyms. We must add a constraint on the intension of *A* and *B* to complete the definition:

$$Sim(\mathrm{A}^\circ_{\mathscr{U}},\mathrm{B}^\circ_{\mathscr{U}}) > \delta \text{ where } \delta \text{ is 'large' (i.e. close to 1).} \tag{23}$$

This constraint is identical to the one used to define near-synonymy.

Note that our definition, which relies on distributions where instances and situations are clearly marked, provides a clear opposition between true synonymy and antonymy.

5 Actual distributions

Having sketched out some properties of ideal distributions with respect to classical lexical semantics relations, we turn to observable data, that is, to those distributions which can be gained from gathering real world utterances. In our account, **actual distributions** correspond to all the utterances that have been perceived by an individual. Like the ideal distributions, actual distributions are based on logical forms for those utterances. They will not refer to neat microworlds, but they do include a notion of the context or situation associated with an utterance. Some of the utterances an individual is exposed to will refer to linguistic entities which are directly perceptually grounded but such grounding is not available in many cases. It is thus obvious that actual distributions will be very different from the ideal distributions which we have been discussing. We nevertheless hypothesize that the utterances that are the basis of the ideal distributions could be produced for a microworld by a native speaker (given enough time!) and that it is possible to produce some approximation to ideal distributions on the basis of actual distributions. That is, while ideal distributions are an abstraction, we assume that the properties we are interested in (inference, modelling of polysemy and so on), could be derived by a language learner on the basis of the actual distributions. Specifically, we assume a) that the learner uses the actual distributions to update their own internal **language model**, b) that this gives the language model some of the properties of the ideal distribution, and c) that the language model would allow a speaker to produce the utterances that the ideal distribution is based on for any given situation. The speaker also has access to probabilistic information derived from actual distributions. The ideal distributions can perhaps be thought of as corresponding to a speaker's semantic competence, while the actual distributions both act as the data source for acquiring competence and provide probabilistic information which could be taken to be an aspect of performance. We assume, for instance, that stylistic expectations (e.g. when to us *policeman/policewoman* vs *cop*) are learnt from the latter.[9]

It is clear that psychologically realistic distributions should correspond to a single person's experience. Unfortunately corpora from which we could derive such distributions in practical experiments are not currently available, except to a very limited extent with child language or artificial contexts. While it may turn out that balanced corpora or even newspaper data can substitute in some experiments for an individuated corpus, this is very unclear, since, as far as we can tell, there is really no empirical evidence that addresses this issue. In fact, there is almost no data on individual adults' exposure to language. We have not even been able to find reliable estimates of how many words someone might be expected to hear/read per day. Our back-of-the-envelope calculations suggest a figure of perhaps 50,000 words per day, which would mean that the British National Corpus, generally regarded as very small by modern standards in computational linguistics, actually corresponds to around 5 years exposure. One consequence is that even words which we might intuitively think of as reasonably familiar to a native speaker are actually encountered relatively infrequently. For instance, *rancid* occurs 77 times in the BNC and *rancorous* only occurs 20 times.[10] This is consistent with our intuitions that individuals use different vocabulary items with very different frequencies and very different contexts, but we do not currently have any way of determining the degree to which this is true. Experiments frequently show large differences between distributions extracted from different corpora, but creating distributions from a very large corpus based on many different genres would lead to differences in use being obscured. Such corpora are, of course, essential for modern lexicography, because they allow the lexicographer to specify the range of meanings of a word in different contexts, explaining uses outside the experience of the dictionary user. However, they do not allow us to model the way in which humans acquire and negotiate meanings.

The second problem is that we have little corpus data available with which we could simulate

---

[9] We will not discuss the possible relationships between our notion of a language model and the way that language works in the human brain here, but we should note that the neural basis of the language model must have some similarities with the notion of a distribution. In particular, the Hebbian learning principle often paraphrased as "Neurons that fire together wire together" is entirely consistent with the idea that frequent relationship between lexemes will lead to strong associations between their functional webs (Pulvermüller 2002).

[10] These counts are from Kilgarriff's web page `http://www.kilgarriff.co.uk/bnc-readme.html`.

grounding. While most utterances perceived by an adult do not directly correspond to perceptual data, we would still like detailed information about the situations which speakers are in to be available as corpus annotation. The only corpora which would (partially) allow for specification of situations are relatively small and are nearly all based on artificial contexts.

A more minor point, but one of considerable practical importance, is that most very large scale corpora contain a considerable proportion of noisy data. For example, a newspaper corpus may contain lists or tables which are not intended to be read in their entirety. Corpora derived from the web are usually much worse in this respect. There is, of course, some vagueness in our notion of an actual distribution in that we have not specified exactly what we mean by 'perception of an utterance', but we intend to exclude cases where the text or speech cannot be understood at all (by an adult).

Thus the corpora in use for distributional semantics within computational linguistics are very different from our notion of an actual distribution. This currently restricts the possibilities for detailed experimentation on the actual/ideal distributions interface. In general, for real investigation of psychologically plausible approaches to distributional semantics, a very large-scale corpus collection effort would be necessary (which we believe would be worthwhile even though the extent to which we could practically simulate grounding would be limited). We are therefore advocating a long-term research program. Nevertheless, we think there are some conclusions to be drawn from theory alone, as we have discussed in §4. Further, we do not exclude that certain types of investigation can be conducted using standard corpora. The fundamental difference between such corpora and individuated data, however, should be born in mind when analysing a system's behaviour.

6 Lexicalised Compositionality and the Generative Lexicon

Having sketched out the notion of actual distribution, we will now turn to the relationship between Lexicalised Compositionality and a well-known approach to lexical semantics: the Generative Lexicon (GL: Pustejovsky 1995). Lexicalised Compositionality shares several of GL's aims and assumptions: in particular, we assume that the lexicon is not just an unstructured list, but that lexical entries are intrinsically interconnected. While a detailed account of the relationship between the approaches would be too lengthy for this paper, here we outline some of the ways in which LC might treat some of the phenomena considered by GL.

The first phenomenon we will consider is regular polysemy. Certain word classes share polysemy patterns, such as, in English, nouns denoting animals also being used for the meat, as mass terms, e.g., *rabbit*, *lamb*, *turkey*, *haddock*. Native speakers readily generate such uses for previously unknown meat types (e.g., *They ate crocodile!*), but in some cases the mass usage is generally blocked by an alternative term (e.g., *cow*, referring to the meat, is blocked by *beef*, *pig* by *pork*). There is a range of evidence that this process is conventionalised: in particular, different languages have somewhat different polysemy patterns. Copestake & Briscoe (1995) developed an account of regular polysemy in terms of lexical rules, which could stand in a hierarchical relationship to one another. For instance, the animal/meat rule is a conventionalised subcase of a general grinding process.

We introduced the idea of spaces in LC distributions corresponding to word senses in §2. This would imply, for instance, that there was some cluster of uses associated with rabbit animals and another cluster associated with rabbit meat in both the ideal and actual distributions. This would also

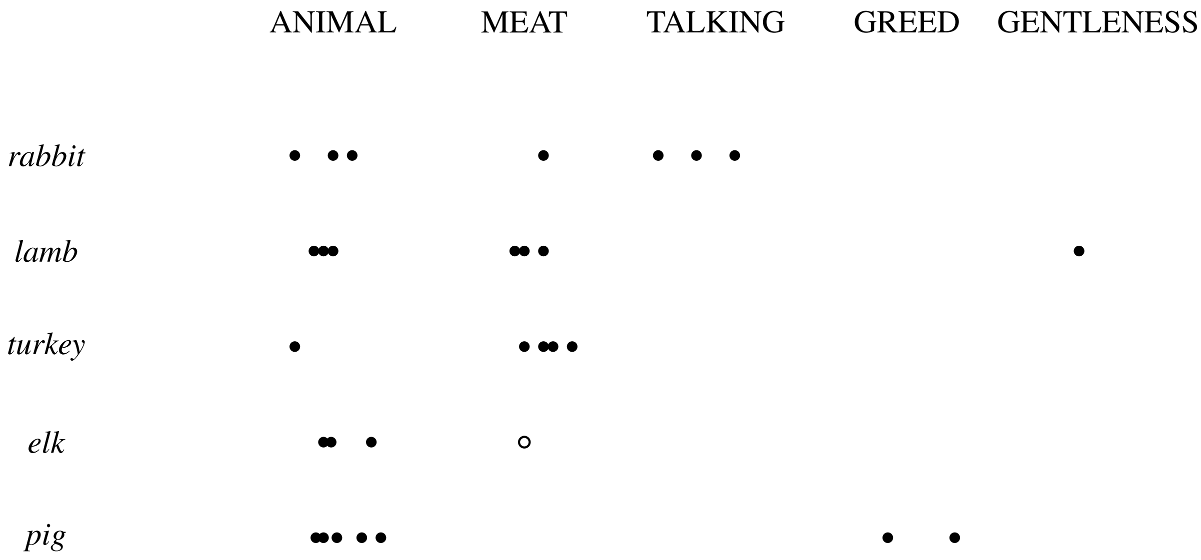

Figure 8 Schematic illustration of the LC account of regular polysemy: solid dots indicate actual uses of lexemes (labelled as ANIMAL etc for the purposes of the figure), open circles indicate unseen but hypothesised uses. Animal and meat uses are found consistently across the class of lexemes, and hence a language learner can hypothesise a regular relationship, but other uses, such as the verb *rabbit* meaning to talk excessively, are idiosyncratic.

apply to *lamb*, *turkey* and so on. The LC account of regular polysemy is essentially that speakers recognise such patterns from the actual distributions and use them when inducing meanings for related words in novel contexts (i.e., when expanding the actual distributions). However, blocking (preemption by synonymy) will occur when there is a well-known term already occupying the relevant meaning space. Figure 8 illustrates this schematically. Thus, on the LC account, there is no enumeration of senses, but lexical count/mass distinctions could nevertheless be said to exist in that there are clusters of uses for a lexeme that are consistent either with count or mass contexts. Lexical rules could be used to capture the interaction with syntax, as in the Copestake and Briscoe account, but they need not be inherently directional.

We now turn to some more subtle meaning distinctions. Words like *book*, which can be viewed as a physical object or as a content-containing entity, have been extensively discussed in GL. Some authors, including Copestake and Briscoe, regard this as a somewhat different phenomenon from regular polysemy, both because there is no syntactic difference between the usages of book and because there are clear cases where both aspects of meaning are invoked with only one mention of an entity, for instance in (24).

(24) Kim is reading a thick red book about syntax

There are, however, contexts in which there is ambiguity: (25) could refer either to works (if Pratchett refers to the famous and prolific author) or physical objects (if Pratchett refers to an occasional user of eBay).

(25) Pratchett sold three books in 2000

But, crucially, (25) has no mixed readings, hence we cannot simply say that *book* is general with respect to these dimensions of meaning. One mechanism available in GL to capture aspects of meaning is qualia structure, whereby lexical entries for nouns include roles corresponding to their form, composition, way of coming into being (agentive role) and their purpose (telic role). It is usual to represent qualia in GL using feature structures. In some versions of the GL account, including Copestake and Briscoe's, the physical object versus contentful entity difference was regarded as involving predicates accessing different parts of the qualia structure, although other versions, including Pustejovsky (2005), utilise dot objects which combine types, e.g., PHYSICAL-OBJECT • INFORMATION. In both cases, the intuition is that *book* can be seen as having multiple meaning components and that the compositional semantics has to ensure that, for example, *read* selects one aspect while *thick* selects another.

In the LC account, the actual distribution book° would contain both predicates that we would expect to pick out physical characteristics (e.g. red°) and predicates relating to its content (e.g., read°), and in cases such as (24), the same linguistic entity is an argument to both types of predicate. This contrasts with cases of homonymy, such as *bank*, discussed in §2.5. On this view, there is no inherent ambiguity between the physical object and information carrier, and the contexts where ambiguity does arise, such as (25), must involve different grounding possibilities, where the linguistic entity can be equated to alternative possible (sets of) real world entities: either physical objects or works. In support of the LC account, we note that a very similar effect also arises with artifacts such as *shirt* or *clock*: it is possible to say, for instance, *That shop sells twenty shirts* with the reading *twenty types/designs of shirt*. But in these cases, it is intuitively clear that there is no necessary difference in real-world individuation between the physical and design aspects of an entity (e.g., a public clock might well be the only clock built to a particular plan) whereas the (modern) canonical use of *book* refers to a conventionally published entity with multiple copies. The LC approach thus gives a somewhat different perspective on the problem, but we leave it as an open question whether dot objects or similar devices would still be necessary to provide a full account of *book*.

Another phenomenon extensively investigated in GL for which an LC account might be useful is logical metonymy, as exemplified by sentences such as *Kim began the cigar*. On the GL account, this can be interpreted (by default) as *Kim began smoking the cigar* because the smoking event is supplied by the telic (purpose) role of *cigar*. Some difficulties with making this approach work are summarised by Copestake (to appear). One problem is that the observed restrictions on logical metonymy are not fully explained by the qualia hypothesis. For instance, the telic interpretation with *begin* generally applies only to consumables and reading material: sentences such as *Kim began the tunnel* are not found with the interpretation *Kim began driving through the tunnel* (as first noted by Godard & Jayez 1993). It seems that this cannot be accounted for by general restrictions on the telic role of *tunnel*, because it is possible to use *after that tunnel* to mean *after driving through that tunnel*, for instance. The second problem is that the qualia values which might be involved in logical metonymy do not appear to be generally usable in accounts of other lexical semantic phenomena. For example, one might hope that qualia would be useful in determining the meaning of compound nominals, but although there is a partial correspondence, many compounds involve relationships which would not be predictable from likely qualia. Another example, discussed in

Copestake (to appear), is the use of adjectives such as *heavy* and *high* meaning 'large magnitude' in examples such as *heavy rain*, *heavy snow*, *high winds*, *high danger* (and not *high rain*, *heavy danger* and so on). Although some fine-grained semantic classes appear to be involved (e.g., *heavy* is used with weather terms denoting some form of precipitation), there is considerable idiosyncrasy, and it does not appear to be possible to develop an account on this basis alone.

In LC, the actual distribution of *cigar* indicates that it is frequently the object of *smoke* and similarly that *smoke* is a plausible argument to *begin*. Hence the metonymic event could be retrieved.[11] This is essentially the approach that Lapata & Lascarides (2003) investigated with corpus data which shows that it is possible to predict the metonymic event in this way with a reasonable degree of accuracy. It is also possible to use distributions to predict the meaning of compound nominals: see, for instance, Turney (2006) and Ó Séaghdha & Copestake (2009). As far as we are aware, no comparable system based on a GL account has been demonstrated.

The GL account is more restrictive than a distributional approach, which could, of course, be an advantage, but it does not seem to be sufficiently flexible to allow for the complexities/messiness of the data. Furthermore, the nature of the fillers of the qualia roles is potentially problematic. If there is a single filler, or a disjunction of a small number of values, it would seem that these would have to correspond to sense-disambiguated concepts. This means the approach depends on making sense distinctions, although it is a primary aim of GL to avoid enumeration of senses. In contrast, in distributional accounts, the relationship is between undisambiguated lexemes. There will be a cluster of usages in smoke$^\circ$ that relate to cigars (as opposed, for instance, to smoked fish), and it is this cluster that contributes to the probability distribution used to predict the metonymic event, but there is no requirement for sense enumeration to achieve this effect. Finally, the idea of qualia is an abstraction over the type of events associated with nouns and, as such, would have to be somehow derived from a language learner's experience, while the LC account is directly based on the actual distributions the learner is exposed to. This implies that GL would need an additional step to be a plausible account of language learning. Of course, proper empirical verification of the LC approach would require the type of individuated corpora we described in §5, but the computational accounts that already exist make us optimistic that this will be possible.

Note that the LC account is only a replacement for the GL treatment with respect to the use of qualia (or other method for representing the detailed make-up of the lexical semantics). It is still necessary to have a representation of the syntax-semantics interface that specifies that *begin* takes an event argument, for instance, and we could adopt this aspect of the GL approach in LC. The LC account can be seen as an alternative to the strictly lexical semantic aspects of GL, but not to the GL accounts of the syntax-semantics interface.

There are some more general points that we can make here about the contrast between feature structures and distributional representations in modelling phenomena. Feature structures are appropriate when we can define a small number of roles that are relevant in a particular context, where the fillers of these roles can be isolated and where processes can be defined which access the filler via the roles. For instance, it makes sense to use feature structures (or dependency structures or trees or description logic), to represent the fact that the subject of the sentence *the dog*

---

[11] The LC approach also allows individual entities to have associated distributions. For instance, if the distribution associated with the particular cigar under consideration is incompatible with it being smoked, then another type of event could be retrieved. This would imply a somewhat different approach to the interface between the lexicon and pragmatics than that described in Lascarides & Copestake (1998). We will not discuss this further here and should emphasize that we would not expect to be able to achieve this practically with any current broad-coverage computational system.

sleeps is *the dog*. It would also make sense to use a feature structure to represent the fact that the numeral classifier *-hiki* is appropriate for *inu* (*dog*) in Japanese, because there are a fixed number of classifiers. In contrast, distributional representations are useful when one has a data source that supports derivation of a distribution and where there is no fixed set of appropriate roles and role fillers. Because there is no predetermined role/filler distinction, it is possible to create abstractions over any concept in distributions, while it is essentially impossible to abstract over roles with feature structure representations. Distributions may also be appropriate as an intermediate representation from which a more abstract feature structure representation can be derived for a particular purpose: this might be part of the process of learning appropriate classifiers, for instance. As discussed above, interfaces between the two types of representation are also necessary to model particular types of processing.[12]

7 Related work

The idea of representing meaning as vectors in a feature space was already proposed in the 1950s in the work of psychologist Osgood (1952), though Harris (1954) is usually cited as the first linguist to express the notion that 'words that appear in similar contexts are semantically similar'. The term 'distributional semantics' came into use by the early 1960s (e.g., Garvin 1962), with Harper (1965) demonstrating what is, to our knowledge, the first actual implementation of the idea and Sparck Jones (1967) first using a principled technique for comparing contexts. Related techniques became widespread in Information Retrieval, but distributional semantics was mostly ignored in computational linguistics until the early 1990s, when reasonably large-scale corpora first became widely available to researchers. The representation of word meanings via distributions has received considerable attention in recent research. Various proposals have been made as to how to choose the most appropriate distributional space to model the semantics of lexical items (Lund & Burgess 1996, Schütze 1998, Landauer & Dumais 1997, Gallant 1998, Griffiths, Steyvers & Tenenbaum 2007, Padó & Lapata 2007). An overview of various methods can be found in Sahlgren (2006) and Turney & Pantel (2010). The setting of the different parameters used in the construction of the feature space is discussed in Bullinaria & Levy (2007).

Distributional techniques have been used extensively to capture various lexical relations. The bulk of the work concerns the extraction of words pairs displaying general similarity (Grefenstette 1994, Turney 2006, Lin & Pantel 2002, Heylen, Peirsman, Geeraerts & Speelman 2008). The general hypothesis for such research is that similarity is a function of the contextual overlap between two words. The more contexts shared, the more similar the two items are. Some research, however, focuses on particular relations: Hearst (1992, 1998) tackles the problem of hyponymy while Girju, Badulescu & Moldovan (2006) investigates the extraction of meronyms and Lin, Zhao, Qin & Zhou (2003), Turney (2008), or again Mohammad, Dorr & Hirst (2008) focus on the identification of antonyms. Work focusing on particular lexical relations tends to be a combination of 'pure' distributional approaches (i.e. modelling lexical items as distributions) and pattern-based heuristics. For instance, the extraction of antonyms might rely on finding out lexical patterns which indicate

---

[12] Note that we do not think it helpful to refer to feature structure representations as symbolic and distributional representations as statistical. While it is usual to associate frequencies or probabilities with distributional representations, it is not necessary to do so: for example, probabilities are only relevant to the lc_0 distribution if we generalise over sets of situations. Similarly, while feature structures etc are often used without probabilities, it is possible to use probabilities in conjunction with feature structures, or (more usually) with rules operating on feature structures.

an antonymy relation. One notable exception is the work of Baroni et al. (2012) which focuses on automatically learning a classifier for the hyponymy relation using word distributions only.

As our approach naturally suggests that composition should be dealt in the traditional way of formal semantics, we should mention proposals which, instead, argue for directly composing lexical items (see Clark & Pulman 2007). That is, while our representation of *black cat* is $\text{black}^\circ \wedge \text{cat}^\circ$, such proposals attempt to build black_cat°. The composition of distributions in phrases is usually performed by 'combining' the vectors of the components of the phrase. Some proposals assume a single composition operation for different types of constructions. Mitchell & Lapata (2010), for instance, experiment with various functions expressed in terms of the two vectors and find that point-wise vector multiplication gives best results in a phrase similarity task, not only for adjective-noun phrases but also noun compounds and verb-noun constructions. Erk & Padó (2008) also adopt a multiplicative approach on sets of vectors involving the selectional preference of the relations associated with a word. Following on such experiments, Guevara (2010, 2011), point out that it is unlikely that many syntactic constructs (e.g. adjective-noun phrases, verb phrases, etc) would be semantically represented by the same operation and argues that, for each construction, it may be possible to learn an appropriate function, representing the effect of one class of words over its arguments. Baroni & Zamparelli (2010) go further, highlighting the potential problems in having a single function for a given grammatical construct. They highlight that different adjective subclasses have different model-theoretic formalisations (Partee 1994) and propose that adjectives are matrices. They express the adjective-noun phrase as an operation of the adjective matrix on the noun vector and learn a different matrix for each adjective in their data. The approaches taken by Widdows (2008) and Grefenstette & Sadrzadeh (2011) are similar.

Our proposal is not antithetical to the direct compositional approach. In fact, we believe that frequent phrases, for instance, may well be stored in the human language model as single items. At this point, however, we prefer to be conservative with regard to which constructions, or specific phrases, should be lexically composed into a single distribution. Note also that, in the ideal distribution representation, different classes of lexical items can be described straightforwardly. For instance, Partee's intersective adjectives are described via the necessary redundancies in the ideal distribution. That is, the presence of the sentence *Kitty is a carnivorous mammal* in the ideal distribution for a particular situation implies that the sentences *Kitty is carnivorous* and *Kitty is a mammal* can also be found in that distribution. By contrast, the sentence *The former president spoke at the meeting* would not normally be accompanied by *The president spoke at the meeting*.

8 Conclusion

We have attempted to give a formalisation of distributional semantics which is compatible with classical formal semantics. Our theory is based on distributions, not sets, but it is translatable into model-theoretic terms. As such, it preserves the idea of extension (we can recover information about which entities are in the world) but it also gives a formal interpretation of a notion of intension by providing structures for lexical items (distributions) that distinguish between their meanings, even when their extensions are identical. One major difference between our account and the standard approaches is that we are assuming speaker-dependent models. An approach centred on the individual seems to us necessary to model language learning, and explain why, for instance, speakers sometimes disagree on the extension of a lexical item (*'This is not a cup, this is a mug!'*).

We introduced the notion of ideal distribution as a theoretical tool for formalisation, but we believe that the concept is also plausible from a psychological point of view – it may be an appropriate description of what we have called the 'language model' of a speaker, i.e. the semantic competence that allows him or her to utter one out of many possible sentences in a certain situation. Our treatment of lexical semantics covers formalisations for standard relations such as hyponymy or antonymy. We also argue that the distributional approach may help to describe some phenomena discussed in the Generative Lexicon theory.

The implementation of the notion of ideal distributions implies recovering 'missing' information from actual distributions. We hypothesised that this process of inference takes place in humans, with constant, radical restructuring of the language model in early learning, and with lesser effects in adult life, the model being updated every time a new concept is learnt or a known concept is used in a yet unobserved way.

To what extent this updating of the language model by actual distributions is reproducible without access to grounded information is an open problem. We have argued that the corpora currently available to computational linguists are very different from the concept of an actual distribution corresponding to an individual speaker's experience, even if we disregard grounding, but it would require a considerable data collection effort to determine whether this was actually the case. We contend that such an effort will ultimately be necessary to develop any psycholinguistically motivated account of distributional semantics. However, our approach does suggest a range of experiments which could be carried out using current corpora. In particular, our approach emphasizes the role of (linguistic) entities in the model, both at a theoretical level and in the contexts of antonymy (§4.4) and sense distinctions (§6). It should be feasible to experiment with distributions which are built from predicates which are applied to the same entity, rather than using a window of words or syntactic dependencies. This might give a motivated way of distinguishing antonyms (unlike standard techniques) and might also give an insight into the aspects of meaning of words like *book* (in contrast with homonyms, such as *bank*). Thus, while this paper is programmatic in nature, we believe that it indicates promising avenues for future experiments in the short term as well as in the longer term.

References

Baroni, M. & R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP10)*, 1183–1193.

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do & Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics (eacl12)*, .

Bullinaria, J.A. & J.P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3). 510–526.

Carey, Susan. 2009. *The Origin of Concepts*. Oxford University Press.

Chierchia, Gennaro. 1998. Reference to kinds across languages. *Natural Language Semantics* 6. 339–405.

Clark, Stephen & Stephen Pulman. 2007. Combining Symbolic and Distributional Models of Meaning. In *Proceedings of the AAAI Spring Symposium on Quantum Interaction*, 52–55. Stanford, CA.

Copestake, Ann. 2009. Slacker semantics : why superficiality , dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL09)* April, 1–9. Athens, Greece.

Copestake, Ann. to appear. The semi-generative lexicon: limits on lexical productivity. In James Pustejovsky, Pierrette Bouillon, Hitoshi Isahara, Kyoko Kanzaki & Chungmin Lee (eds.), *Recent Trends in Generative Lexicon Theory*, Springer.

Copestake, Ann & Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics* 12:1. 15–67.

Copestake, Ann, Dan Flickinger, Ivan A Sag & Carl Pollard. 2005. Minimal Recursion Semantics: an Introduction. *Journal of Research on Language and Computation* 3(2-3). 281–332.

Erk, Katrin & Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI.

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1). 15–28.

Gallant, Stephen I. 1998. Context Vectors: A Step Toward a 'Grand Unified Representation'. In *Hybrid Neural Systems*, 204–210. London, UK: Springer-Verlag.

Ganesalingam, Mohan. 2009. *The Language of Mathematics*: University of Cambridge dissertation.

Garvin, Paul L. 1962. Computer Participation in Linguistic Research. *Language* 38(4). 385–389.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford, England, UK: Oxford University Press.

Girju, R., A. Badulescu & D. Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics* 32(1). 83–135.

Godard, Danièle & Jacques Jayez. 1993. Towards a proper treatment of coercion phenomena. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL93)*, 168–177.

Grefenstette, E. & M. Sadrzadeh. 2011. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP11)*, 1394–1404. Edinburgh, Scotland, UK.

Grefenstette, G. 1994. *Explorations in automatic thesaurus discovery*. Springer.

Griffiths, T.L., M. Steyvers & J.B. Tenenbaum. 2007. Topics in semantic representation. *Psychological review* 114(2). 211.

Guevara, Emiliano. 2010. A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics (ACL 2010)* 33–37.

Guevara, Emiliano. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, 135–144. Oxford, England, UK.

Harper, Kenneth E. 1965. Measurement of similarity between nouns. In *Proceedings of the 1st International Conference on Computational Linguistics (COLING65)*, 1–23. New York, NY.

Harris, Zelig. 1954. Distributional Structure. *Word* 10(2-3). 146–162.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING92)*, 539–545. Nantes, France.

Hearst, Marti A. 1998. Automated discovery of WordNet relations. In *WordNet: an electronic lexical database*, 131–151. MIT Press.

Heylen, K., Y. Peirsman, D. Geeraerts & D. Speelman. 2008. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 3243–3249.

Hobbs, Jerry. 1985. Ontological promiscuity. In *Proceedings of the 23rd Conference of the Association for Computational Linguistics (ACL 1985)*, 61–69. Chicago, IL.

Landauer, Thomas K & Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 211–240.

Lapata, Mirella & Alex Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics* 29(2). 261–315.

Lascarides, Alex & Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics* 34. 387–414.

Lehrer, Adrienne. 2002. Paradigmatic relations of exclusion and opposition I: Gradable antonymy and complementarity. In D. A. Cruse, F. Hundsnurscher, M. Job & P.-R. Lutzeier (eds.), *Lexicology: An international handbook on the nature and structure of words and vocabularies*, 498–507. Berlin: De Gruyter.

Lin, Dekang & Patrick Pantel. 2002. Concept Discovery from Text. In *Proceedings of the 19th international conference on Computational linguistics (COLING02)*, 577–583. Taipei, Taiwan.

Lin, Dekang, Shaojun Zhao, Lijuan Qin & Ming Zhou. 2003. Identifying Synonyms among Distributionally Similar Words. In *Proceedings of the eighteenth international joint conference on artificial intelligence* 4, 1492–1493. Acapulco, Mexico.

Link, Godehard. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In Rainer Bäuerle, Christoph Schwarze & Arnim Von Stechow (eds.), *Meaning Use and Interpretation of Language*, 302–323. Walter de Gruyter.

Lund, Kevin & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28. 203–208.

Lyons, John. 1977. *Semantics (v1)*. Cambridge, England, UK: Cambridge University Press.

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

Mitchell, Jeff & Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34(8). 1388–1429.

Mohammad, Saif, Bonnie Dorr & Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the conference on empirical methods in natural language (emnlp08)* EMNLP '08, 982–991. Waikiki, Honolulu, Hawaii: Association for Computational Linguistics.

Ó Séaghdha, Diarmuid & Ann Copestake. 2009. Using Lexical and Relational Similarity to Classify Semantic Relations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL09)* April, 621–629. Athens, Greece.

Osgood, Charles E. 1952. The Nature and Measurement of Meaning. *Psychological Bulletin* 49(3). 197–237.

Padó, Sebastian & Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics* 33(2). 161–199.

Partee, B.H. 1994. Lexical semantics and compositionality. In Daniel Osherson, Lila Gleitman & Mark Liberman (eds.), *Invitation to cognitive science, second edition. part i: Language*, vol. 1, 311–360. MIT Press.

Pulvermüller, Friedemann. 2002. *The Neuroscience of Language*. Cambridge University Press.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press.

Rubenstein, Herbert & John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10). 627–633.

Sahlgren, M. 2006. *The Word-space model*: Stockholm University dissertation.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.

Sparck Jones, Karen. 1967. A small semantic classification experiment using cooccurrence data. Tech. Rep. ML 196 Cambridge Language Research Unit, Cambridge.

Thater, S., H. Fürstenau & M. Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th international joint conference on natural language processing*, Chiang Mai, Thailand.

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics* 32(3). 379–416.

Turney, Peter D. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING08)*, 905–912. Manchester, UK.

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

Widdows, Dominic. 2008. Semantic Vector Products : Some Initial Investigations. In *Second AAAI Symposium on Quantum Interaction* March, Oxford, England, UK.

Ann Copestake
University of Cambridge
Computer Laboratory
15 J.J. Thomson Avenue
Cambridge CB3 0FD, UK
ann.copestake@cl.cam.ac.uk

Aurelie Herbelot
Universität Potsdam
Department Linguistik
Karl-Liebknecht-Straße 24-25
D-14476 Golm, Germany
aurelie.herbelot@cantab.net