Lexicalised compositionality

Ann Copestake
University of Cambridge

Aurelie Herbelot
Universität Potsdam

Abstract In this paper, we propose an approach to distributional semantics which can be formally related to a simple model-theoretic approach. We describe treatments of some of the traditional lexical semantic relationships within this framework, and also outline accounts of some phenomena which have been considered within Generative Lexicon theory. We further argue that distributions should be based on individual experience, rather than the type of text corpora currently used in computational linguistics and lexicography.

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

1 Introduction

This paper explores a possible approach to the integration of distributional semantics with compositional semantics. The essential idea is to work with distributions rather than models 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. We will explore this idea with a particular notion of 'context'. We are primarily interested here in exploring a theoretical account, although our goal is to develop an approach which is eventually usable for computational linguistics and which might be elaborated into a psycholinguistically plausible model. Our aim is an account which could be based on empirically observable data of what an individual actually hears. In this way, we can investigate what we could, in principle, get out of such data and where additional information sources would be necessary.

We assume that a full account of semantics should support compositionality and inference, as is generally accepted in formal semantics. However, it should also provide a way of representing lexical meaning, including a non-stipulative approach to word senses and regular polysemy. In our view, an ideal approach to semantics should support underspecification: no distinction should be required by the semantic representation unless there is a clear correspondance to a morphological or syntactic distinction. For instance, the motivation for using a semantic representation that allows for quantifier scope underspecification is that there is generally no evidence for making a syntactic distinction between the different scope possibilities. We also 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.

Our approach to lexical semantics is essentially distributional: i.e., based on the contexts in which a word occurs. Computational linguists now often use distributional representations of lexical meaning, though human learnability of models is not a mainstream concern. However, most recent work in computational linguistics tends to ignore or downplay the achievements of formal semantics.

This seems misguided: notions such as generalised quantifiers are too important to be missed out of any approach which attempts a full account of semantics. Conversely, distributional semantics has received little attention within mainstream linguistics, but we believe it has considerable potential as a theoretical account as well as a practical technique. One of the main aims of this paper is thus to see if we can link current practice in distributional approaches in computational linguistics to formal semantics and to lexical semantics.

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.

To illustrate the aims a little further, consider the following sentence:

(1) Universities in England will see class sizes balloon.

We wish to be able to relate this to the paraphrase:

(2) English universities will see the size of their classes expand.

Progress in computational linguistics makes it realistic to aim to achieve relatively constrained paraphrases of this type on arbitrary text (with some degree of reliability) on the basis of information that can be acquired automatically from distributions of words in corpora, without solving the general AI problems of the representation of world knowledge and without intractable inference. We should note here that it is irrelevant whether we think of distributional information as 'really' world knowledge or not, but we would argue that phenomena such as the systematic polysemy exemplified here by *balloon* the noun and *balloon* the verb are part of linguistics (for instance, because there are differences between languages).

The hypothesis to be investigated here is that instead of talking about the set of all things in the world denoted by university′, as in an extensional account, or using a Fregean notion of sense, we talk about the **context set** for university. In §2, we will introduce the idea of an **ideal distribution**, where we consider all the contexts in which *university* could occur. Each context corresponds to the logical form of a sentence/utterance. For example, contexts where *university* is the subject of *see* will be a subset of all the contexts where *university* occurs in subject position, which in turn will be a subset of all the contexts in which *university* occurs. We will see in §2 how we can relate this to more standard ideas about denotation.

It is clear that 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° 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 our approach, distributional context sets 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.

The objective of this paper is firstly to argue that building a distributional account of sense has some merit and secondly to 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. In section §3, 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. In §4 and §5, we outline how various phenomena in semantics might be analysed in our approach and in §6 we provide a very 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.[1]

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

1 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

in S_1 are:

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

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

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.[2] 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). 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.[3] MRS representations may be underspecified for certain ambiguities which are not resolved by syntax, such as scope

2 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.

3 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.

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

Figure 2 Ideal context sets for S_1

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 names for 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 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)[4]

We define the context set for a lexeme *l* in terms of the logical forms which contain an elementary predication corresponding to *l*. [5] 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

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

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

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

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):

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

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

5 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

Thus, for a predicate P, 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 $P'(z)$ is true, we have a logical form for a sentence in the lc_0 distribution containing an elementary predication equivalent to $P^\circ(z)$. 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 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 situation entities and events.
Correct assignment:

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

Incorrect assignment:

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

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 situation entities can thus proceed via comparison without any sort of explicit meaning

$$
\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 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 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 §6, we do not need these for our theoretical account of meaning for compositional phrases.[6]

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. We can also see how 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, we must process quantified statements before adding them to the ideal distribution. We discuss the matter further in Sections 4.3 (on hyponymy) and 4.4 (on quantifiers).

2.3 Linguistic entities and situation entities

The level of indirection provided by distinguishing between linguistic entities and real world entities, advocated by Hobbs (1985), has a number of advantages from our viewpoint. In fact, although we sometimes loosely use the term 'real world entities' instead of 'situation entities', 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. 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

6 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 are known 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$^\circ$. Again we are essentially assuming that the LFs are constructed by a system similar to the English Resource Grammar which has lexical entries corresponding to such MWEs.

	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.

linguistic entities and situation entities is more than just a convenient computational linguistics hack.

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 are a large range of approaches 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 $\text{table}'(x)$. On the other hand, we might only be interested in predications which directly relate to an entity corresponding to the word

	$a(x)$	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

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 up to a certain length, or only include paths of a particular type (cf Padó & Lapata 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$^\circ(e,x)$). Some reasons why the latter move might be sensible are discussed in §4.

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 §4.

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. In §4, we will return to the issue of generalising over context sets and calculating lexical semantic relationships such as similarity on that basis.

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. In general, clear cases of homonymy, such as the bank example, give rise to relatively discrete clusters (see, e.g., Schütze 1998, Lin & Pantel 2002).

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 §5, 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 situation entities: one referred to by the constants $x3$ and $x4$, which correspond to the entity we called c, and one referred to 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.[7] 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

7 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.

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 the next section.

3 Actual distributions

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 that the learner uses the actual distributions to update their own internal **language model**, that this gives the language model some of the properties of the ideal distribution, and 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 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 associated functional webs (Pulvermüller 2002).

In this section, we will outline some of the issues involved in very general terms, moving on to more specific discussion in §4.

3.1 Individuated, situation-annotated corpora

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.[8] 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 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. They are consequently rather unsuitable for detailed investigation of the LC idea. 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 current approaches, as we will discuss in the context of various lexical semantic phenomena.

3.2 Approximating ideal distributions from actual distributions

There is no neat correspondence between the distributions we can extract from corpora and lexical semantic relationships such as synonymy, antonymy and hyponymy. In current computational work, similarity of distributions is somewhat related to synonymy, but true synonyms are not more similar than near-synonyms. Antonyms and synonyms cannot easily be distinguished. Hyponymy can only be established indirectly. We would also not expect actual distributions experienced by humans to directly correspond to the traditional lexical semantic relations or to standard notions of extension.

8 These counts are from Kilgarriff’s web page `http://www.kilgarriff.co.uk/bnc-readme.html`.

In order to establish such relationships, we need to consider how the ideal distributions might be approximated on the basis of actual distributions.

The most obvious problem with approximating ideal distributions on the basis of what is actually uttered is that even when we consider utterances that pertain to a quite specific situation, they will still be extremely limited compared to our idealisation of all the possible utterances. We are thus primarily looking for ways to expand the actual distributions. Inference rules should, in principle, allow us to do this. For instance, once we know that all cats are mammals, we could take sentences involving *cat* and construct analogous mammal sentences. Of course, we do not expect to generate the full ideal distributions in any actual computational system, but we can use the notion of the ideal distribution as a means of validating techniques. We will discuss this in more detail in §4.6.

A secondary problem is that, although ideal distributions would be logically consistent (to the extent that an individual speaker is capable of reasoning), we cannot expect logical consistency in the actual distributions. We might expect a individual to be exposed both to ‘The Loch Ness Monster exists’ and ‘The Loch Ness Monster does not exist’, for instance. We hypothesise that such inconsistencies would have little effect on the lexical semantic representation, however: for example, a speaker’s use of *Loch Ness Monster*, as reflected in the distributions, is not greatly affected by whether that speaker believes in its existence or not. This speculation would, of course, require empirical verification.

4 Lexical semantics and ideal distributions

In this section, we will describe how standard relations in lexical semantics can be formally expressed using distributions. For this purpose, we assume a theoretical setup where ideal distributions correspond to the sum of all linguistic experiences associated with a particular speaker. Where appropriate, the standard linguistic definitions will be given, drawing extensively on Cruse (1986) and Geeraerts (2010).

In what follows, distributions are assumed to have been ‘partitioned’ into appropriate subspaces (see §2.5). So when we talk of cube$^\circ$, we talk of the distribution of a particular subspace, or ‘sense’ in the classical account, of *cube* – which subspace should be obvious from the context. Recall, also, that we regard fixed expressions as words with spaces, which have separate distributions from their components. We assume, for example, that in 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

We define the following two notions:

- The shared distribution of two lexical items:

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

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

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

We can give numerical values corresponding to these relations. $S_n(\mathrm{A}^\circ, \mathrm{B}^\circ)$ would express the degree to which A and B share context. Such values can be computed in a variety of ways: the simplest approach is the Jaccard metric:

(5)
$$S_n(\mathrm{A}^\circ, \mathrm{B}^\circ) = \frac{|\mathrm{A}^\circ \cap \mathrm{B}^\circ|}{|\mathrm{A}^\circ \cup \mathrm{B}^\circ|}$$

Similarly,

(6)
$$C_n(\mathrm{A}^\circ / \mathrm{B}^\circ) = \frac{|\mathrm{A}^\circ - (\mathrm{A}^\circ \cap \mathrm{B}^\circ)|}{|\mathrm{A}^\circ \cup \mathrm{B}^\circ|}$$

We follow Harris (1954) in his claim that lexical items that appear in the same type of contexts are semantically similar. According to this view, it may be tempting to directly express semantic similarity as our notion of shared distribution and write that the similarity $Sim(A,B)$ of A and B is simply $S_n(\mathrm{A}^\circ, \mathrm{B}^\circ)$. However, due to the nature of our distributions, which include specific information about instances and situations, this definition of similarity would not be equivalent to the one intended by Harris, or the one used by the computational linguistics community in recent years. We may, for example, consider the concepts of *cat* and *dog* fairly similar, but they are never substitutable in any given existentially quantified context: I cannot point to a cat and say *This is a dog*. In fact, their shared distribution $S(\mathrm{cat}^\circ, \mathrm{dog}^\circ)$ is 0. We will show throughout this section that we need two slightly different notions of distribution to formally define similarity on one hand and certain relations such as synonymy or hyponymy on the other hand. Having already introduced our base contexts sets in Section 2, we will next explain the derived idea of **generalised context sets**.

4.1.1 Generalised context sets and distributional similarity

Context sets include information about instances and situations but it should be clear that by reducing a context set to a representation that includes logical forms only, we get a derived distributional form more akin to the linguistic objects typically assumed by computational linguists. In what follows, we describe how to perform this reduction.

Some generalisations can be made over the logical forms included in an ideal context set. Consider the following three contexts in the distribution of *cat* and assume $x1 =_{rw} x11$:

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

It is possible to generalise over the situations where the cat referred to by $x1$ sleeps by writing:

$$< [x1][a(x1), \mathrm{sleep}^\circ(E, x1)], S >$$

where E is the set of events $\{e1, e2\}$ and S is the situation set $\{S_1, S_2\}$. We also define a correspondence set *Corr* which spells out how events entities and situations are linked in the world under consideration. For our example, $Corr = \{(e1, S_1), (e2, S_2)\}$.

Similarly, we can generalise over all situations where any cat is sleeping by writing:

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

where X is the set of entities $\{x1, x2\}$, $E = \{e1, e2, e3\}$, $S = \{S_1, S_2, S_3\}$ and $Corr = \{(x1, e1, S_1),$ $(x1, e2, S_2)$, $(x2, e3, S_3)\}$.

We will call this type of expression a **generalised context set** and use the subscript $_G$ to denote it: cat°_G. A special underspecified form of a generalised context set, where X, E and S are unknown and a logical form is duplicated for each item in its *Corr*, is the form used in classical distributional similarity calculations: it provides frequencies for all context types observed in a given corpus, but without reference to specific situations (as in Fig 6). The same form without duplication is equivalent to a simple binary distribution which indicates whether a particular context appears or not in the corpus (as in Fig 5). We will call those special forms **underspecified generalised context sets** and use the subscript $_{UG}$ to denote them: cat°_{UG}. (We will only specify the exact nature – binary or frequency-based – when the distinction matters.)

We can then define similarity as the shared distribution between two underspecified generalised context sets:

(7) $$Sim(A,B) = S(\text{A}^\circ_{UG}, \text{B}^\circ_{UG})$$

This can be quantified by standard methods: e.g., cosine or pointwise mutual information.

4.2 Synonymy

Synonymy can be defined via the idea of 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. 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 (see, for instance, *off* and *rancid*, where the latter is only applicable to fatty food) Sometimes, also, words are definitionally substitutable but they present a difference in meaning which is more stylistic or emotive, as noted by Geeraerts (2010) who contrasts *prostitute* and *whore*. 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. 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 generalised 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 distributions.

If A and B are synonyms, then

(8) $$\text{A}^\circ = \text{B}^\circ$$

By extension,

(9) $$S(A^\circ, B^\circ) = A^\circ = B^\circ$$

(10) $$C(A^\circ/B^\circ) = C(B^\circ/A^\circ)$$

(11) $$S_n(A^\circ, B^\circ) = 1$$

(12) $$C_n(A^\circ/B^\circ) = C_n(B^\circ/A^\circ) = 0$$

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.

4.2.2 Near-synonyms

Near-synonymy is a phenomenon more related to similarity than to synonymy itself. Therefore, we define it using generalised context sets.

If A and B are near-synonyms, then

(13) $$S_n(A^\circ_{UG}, B^\circ_{UG}) > \delta \text{ where } \delta \text{ is 'large' (i.e. close to 1).}$$

4.3 Hyponymy

Hyponymy or **hyperonymy** are 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 the ideal distribution, 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). More generally, if A is a hyponym of B:

(14) $$A^\circ \subset B^\circ$$

We reject the essentialist view which would allow us to have an inclusion relation between hyponym and hypernym at the generalised context set level. However, universal statements about hypernyms can straightforwardly be applied to their hyponyms. So we can write:

(15) $$\text{if } A^\circ \subset B^\circ \text{and } < [b][all(b), P^\circ(e,b)] > \in B^\circ \text{then } < [a][all(a), P^\circ(e,a)] > \in A^\circ$$

Note that the inclusion relation in distributions implies an idea of substitutability, as described in Dagan, Glickman, Gliozzo, Marmorshtein & Strapparava (2006). When a certain logical form is found in two distributions, the lexical items corresponding to those distributions can be said to be in a (normally unidirectional) substitutability relation with regard to that logical form. For instance,

(16) Kim owns a bike.
$\text{bike}'(b) \wedge \text{own}'(Kim, b)$

entails

(17) Kim owns an object.
$\text{object}'(b) \wedge \text{own}'(Kim, b)$

That is, we can straightforwardly substitute object′ for bike′ in the logical form, without affecting the truth of the model.

In the next subsection, we will talk about the assumptions we need to make in order to preserve the inclusion relation between hyponyms and hypernyms. In particular, we will discuss how to deal with **quantification**.

4.4 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. Note the problem in having pluralised statements in distributions:

$$< [x][some(x), \text{black}^\circ(x)] > \subset \text{cat}^\circ$$
$$< [x][some(x), \text{black}^\circ(x)] > \subset \text{animal}^\circ$$

(Some cats are black and some animals are black.)

$$< [x][all(x), \text{mammal}^\circ(x)] > \subset \text{cat}^\circ$$
$$< [x][all(x), \text{mammal}^\circ(x)] > \not\subset \text{animal}^\circ$$

(All cats are mammals but not all animals are mammals.)

Only quantifiers that are upward monotone give the correct inclusion relation between hypernym and hyponym.

In order to have a representation that behaves in the same way for all quantified logical forms, we propose that quantifiers must be **unpacked** before inclusion of a logical form in a distribution. We define the process of unpacking as the translation of a logical form containing plurally quantified arguments into several logical forms, one for each element in the set denoted by the quantified argument:

In cat°:

$$< [x1][three(x1), \text{sleep}^\circ(e1, x1)], S_1 > = \{ \; < [x11][one(x11), \text{sleep}^\circ(e11, x11)], S_1 >, \; < [x12][one(x12), \text{sleep}^\circ(e12, x12)], S_1 >, \; < [x13][one(x13), \text{sleep}^\circ(e13, x13)], S_1 > \}$$

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. It simply consists in writing (and verifying) the relevant distributional equality between a plurally quantified 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:

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

where $x1 = [x11x12x13x14]$.

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

Note that this account results in the resolution of generalised quantifiers into sets of simple variables. In some cases, full resolution is not possible. See the following:

(18) Kim wants a bike.

Assuming a reading of the sentence where Kim wants *any* bike, we have a case where no particular instance in the model can be associated with the object noun phrase in the sentence. We will say that the object noun phrase in the sentence is non-specific (for the purpose of this work, we will assume the notion of **specificity** to refer to cases where the instance talked about in the sentence can be identified in the model under consideration). Given our (slacker semantics) assumptions, this could be expressed as:

(19) $$\exists b[\text{bike}'(b) \wedge \text{want}'(Kim, b)]$$

In order to perform unpacking in this type of situation (that is, whenever a non-specific entity is involved), we explicitely interpret the existential quantifier as an OR relation:

(20) $$\begin{gathered} \text{given } X = x_1, x_2 ... x_n \\ \text{if } \exists x[\text{X}'(x)] \\ \text{then } x = x_1 \vee x_2 ... \vee x_n \end{gathered}$$

In doing so, we preserve the consistency of the model.

We can then express 18 as follows (the star notation $\sigma^* x$ is from Link (1983) and represents the supremum, or maximum plurality of the individuals with a certain property):

(21) $$X = \sigma^* x\text{:bike}'(x) = x_1, x_2 ... x_n \wedge \exists b[b = x_1 \vee x_2 ... \vee x_n \wedge \text{want}'(Kim, b)]$$

That is, X is the set of all bikes ($X = x_1, x_2 ... x_n$) and Kim wants a bike b such that $b = x_1 \vee x_2 ... \vee x_n$.

As far as genericity is concerned, we take the view argued for in Herbelot & Copestake (2011) that generics express quantification, at least at one level of their semantics. In the ideal distribution, the underspecification can be easily resolved into a particular quantifier which will, in turn, be unpacked in the way described above.

Introducing unpacking means that quantifiers are never encountered in distributional logical forms and that they must be considered linguistic meta-data – descriptors of distributions. In that view, they are a relational phenomenon in the same way as synonymy and hyponymy themselves.

4.5 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. This class includes phenomena of 'perspectival opposition', as observed in the pair *buy/sell*, where one of the terms entails the other one as well as a change of perspective in the predication: *A buys B from C* entails *C sells B from A*, for instance. The last class, multiple antonyms, refers to terms that denote several discrete points on a non-gradable, discontinuous scale: 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*) and secondly, antonyms revolve around a certain concept (temperature, life and academic career in the examples above) and are therefore related in terms of intension.

In our terms, if *A* and *B* are antonyms, then

(22) $$S(\mathrm{A}^\circ, \mathrm{B}^\circ) = 0$$

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

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

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

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

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

4.6 Lexical semantics and actual distributions

At the end of §2, we introduced the idea that actual distributions could be seen as updating the language model of a given speaker. We also proposed to use ideal distributions to represent the language model.[9] In this section, we will show that the assumptions we make about the updating effects of actual distributions influence the way we should express lexical relations in such a model.

9 For the sake of conciseness, we will talk in this section of distributions 'in' the language model, but we wish to make clear that we only see distributions as an abstract representation and not as 'the language model itself'.

We will also come back to quantification issues to illustrate the potential problems that implementing the updating function might cause in the 'real world'.

4.6.1 Effects of actual distributions on ideal distributions

The following scenarios describe situations where synonymy is at work.

i. I am told that *eggplant* is the American English equivalent of *aubergine*. I start using the new term when addressing American speakers and I understand it as its British equivalent when encountered in text. I assume the distributions of both terms to be identical.

ii. a. Having encountered the term *ineffable* in text, I look it up in the dictionary and find it glossed as *unspeakable*. In the absence of further information, I assume the distribution of *ineffable* to be identical to that of *unspeakable*. After time, I realise that *ineffable* is linked to philosophical or spiritual contexts and amend its distribution accordingly (for instance, I erase *crime* as potential argument).

    b. Having encountered the term *rancid* in an unmistakable context, I make the assumption that it means *off*. I assume the distribution of the terms to be identical until I realise that *rancid* can only be used in relation to fatty food. At that point, I amend the distribution of *rancid* accordingly.

iii. I know that *hard* and *difficult* are synonyms. I consider both *hard work* and *difficult work* acceptable utterances but in my own speech, I overwhelmingly use the former (cf. Calude & Pagel 2010).

Those scenarios illustrate several points. First, synonymy, whether observed or definitional, is one of the phenomena that have a direct impact on the constitution of the language model. Second, there seems to be no storing of a synonymy 'status' in the language model. If that was the case, any update to the *rancid* distribution after it was first assumed to be a true synonym of *off* (see iib) would be mirrored in the distribution of *off* and the semantic difference between the two terms would never be acknowledged. Third, the nature of the update depends on the nature of the observation made in actual distributions: a single linguistic phenomenon (synonymy) is conducive to several updating functions depending on how it has been observed (contrast i and iia). Fourth, frequency effects in actual distributions are stored and reproduced by speakers (see iii). Finally, synonymy tells us something about the notion of meaning.

Given the way we have described true synonyms and near-synonyms so far, we could hypothesise that the main difference between the two groups is that true synonyms involve complete distributional equality in the language model while near-synonymy only presupposes a high overlap between distributions, whether actual or ideal. For the updating of the language model to function in the way that we observed in Examples i, iia and iib, we must however assume a more fundamental difference between true synonyms and near-synonyms.

True synonymy is in some way encoded in the language model while, as argued above, near-synonymy cannot be, as it would prevent the expected updating of the distributions. We hypothesise the existence of distributional **identity** versus distributional **equality**. The former is established when true synonymy is ascertained once and for all (for example, after encountering expert knowledge of the type *'Eggplant' is the American English equivalent of 'aubergine'*). It assumes the existence of one distribution only, linked to two **names**. Distributional equality, in contrast, maintains two different distributions for the two terms in a relation of near-synonymy or potential true synonymy.

The updating effect of the actual distribution on the language model for hyponymy follows roughly the same principles as the ones observed for synonymy. We will assume, for instance, that a 'hyponymy' status can be stored for a pair of distributions, say cat$^\circ$ and feline$^\circ$, if the sentence *Cats are felines* is encountered in a context where it can be taken as expert knowledge. We will also assume that hyponymy can be 'discovered', i.e. inferred from the current state of an individual's language model, and potentially revised in the light of new knowledge.

4.6.2 Implementation issues

As pointed out in §2, we assume that it is in theory possible to 'recover' ideal distributions from the partial information given by actual distributions. This is, after all, what happens as humans learn their mother tongue. Implementing this process, however, is far from trivial. As an example of the issues we might encounter, we discuss next the unpacking operation.

Unpacking quantifiers in the actual distribution is a more complex process than in the ideal distribution. We must take the following into account:

- in most cases, the exact cardinality of the quantifier is unknown; e.g. we do not know how many cats the sentence *Some cats sleep* refers to.
- the real-world equalities between individuals denoted by distributional arguments are not necessarily known. This means that, as we assume that quantification is a relation which, like synonymy or hyponymy, will be 'discovered' in a certain state of the ideal distribution, we need those equalities in order to obtain the correct quantification.

Assume the following three sentences in the actual distribution, and their logical forms:

$$\begin{array}{lll} \text{Two cats play.} & \{ & <[one(x1), \text{cat}^\circ(x1), \text{play}^\circ(e1,x1)]>, \\ & & <[one(x2), \text{cat}^\circ(x2), \text{play}^\circ(e2,x2)]>\} \\ \text{A cat sleeps.} & \{ & <[one(x3), \text{cat}^\circ x3, \text{sleep}^\circ(e3,x3)]>\} \\ \text{A cat sleeps.} & \{ & <[one(x4), \text{cat}^\circ x4, \text{sleep}^\circ(e4,x4)]>\} \end{array}$$

The corresponding distribution for cat$^\circ$ is:

$$\begin{array}{lll} \text{cat}^\circ = & \{ & <[x1][one(x1), \text{play}^\circ(e1,x1)]>, \\ & & <[x2][one(x2), \text{play}^\circ(e2,x2)]> \\ & & <[x3][one(x3), \text{sleep}^\circ(e3,x3)]> \\ & & <[x4][one(x4), \text{sleep}^\circ(e4,x4)]>\} \end{array}$$

The equalities $x3 =_{rw} x1$ and $x4 =_{rw} x2$ are necessary in order to make the inference *All cats sleep*. But the sentences on their own do not provide those equalities: grounding information must be present. We therefore expect that straightforward unpacking can only be achieved for universally quantified statements and for constrained situations where coreference resolution is possible.

Despite the type of difficulties just outlined, we believe that it is possible to go a long way in automatically updating the implementation of a language model if encyclopaedic knowledge is used appropriately. Automatically processing generics and universals seems a natural first step in capturing hyponymy relations, for instance, and one that would allow further updating from actual distributions. Having recorded that *All cats are mammals*, we can make sure that every new instance x introduced in the cat$^\circ$ distribution is accompanied by the logical form mammal$^\circ(x)$.

5 Lexicalised Compositionality and the Generative Lexicon

In this section, we will turn to the relationship between Lexicalised Compositionality and the Generative Lexicon (GL: Pustejovsky 1995). GL is a well-known approach to lexical semantics, on which one of us has worked extensively. 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 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 (27).

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

There are, however, contexts in which there is ambiguity: (28) could refer either to works (if

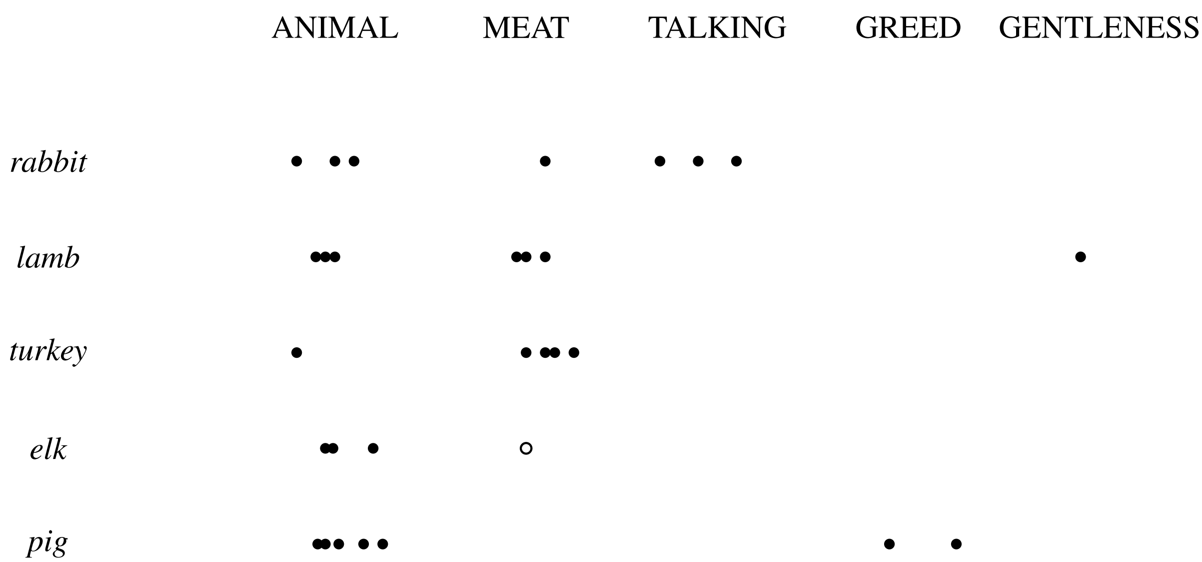

Figure 8 Schematic illustration of the LC account of regular polysemy: solid dots indicate actual uses of lexemes (labelled as ANIMAL etc for the purposes of the figure), open circles indicate unseen but hypothesised uses. Animal and meat uses are found consistently across the class of lexemes, and hence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

Pratchett refers to the famous and prolific author) or physical objects (if Pratchett refers to an occasional user of eBay).

(28) Pratchett sold three books in 2000

But, crucially, (28) 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$^\circ$ would contain both predicates that we would expect to pick out physical characteristics (e.g. red$^\circ$) and predicates relating to its content (e.g., read$^\circ$), and in cases such as (27), 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 (28), 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.[10] 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 §3.1, 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, 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

10 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.

structures or trees or description logic), to represent the fact that the subject of the sentence *the dog 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.[11]

6 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 Turney (2008) the identification of antonyms. The extraction of entailment rules, as in Szpektor, Tanev & Dagan (2004), is a problem closely related to the identification of hyponymy – at least in cases where single lexical items (including words with spaces) are extracted.

11 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.

More recently, it has been suggested that in order to integrate distributional semantics with model theoretic formalisms, methods should be found to compose the distributions of single words (Clark & Pulman 2007). It is clear that the representation of *carnivorous mammal* in formal semantics can be written as $\text{carnivorous}'(x) \wedge \text{mammal}'(x)$ but it is less clear how the lexical semantics of the phrase should be described in distributional terms. The composition of distributions in adjective-noun phrases is usually performed by 'combining' the vectors of the components of the phrase. 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.

There are some theoretical reasons to regard multiplication as an appropriate model for composition, as it corresponds to the idea of intersection in formal semantics, ported to distributional semantics. Guevara (2010, 2011), however, points out that it is unlikely that many syntactic constructs 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. Accordingly, he experiments with models based on addition, multiplication, circular convolution and partial least squares regression. His experiments show that for adjective-noun pairs, the partial least squares regression model performs best, while the additive model gives better results on verb-noun pairs.

There are also potential problems with having a single function for a given grammatical construct. Partee (1994), writing on lexical semantics and compositionality from the point of view of formal semantics, shows that it is not possible to give a unified semantics for adjectives. She gives a classification of adjectives based on four subclasses: intersective, subsective, privative and 'plain' non-subsective and demonstrates that each subclass has a different model-theoretic formalisation. We can thus conclude that composing a noun with those different classes of adjectives should also have various consequences for the distribution of the resulting phrase. For instance, the distribution of *fake gun* should have low values for the dimensions related to the destructive capability of a real gun. Baroni & Zamparelli (2010) acknowledge this issue 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.

In contrast with some of the work described above, our goal is not application-oriented – although we hope it can be a basis for the implementation of theoretically motivated distributional models. Our aim is to provide distributional formalisations for lexical semantics, and to integrate them into classical formal semantics. Key to this endeavour is the notion of ideal distribution, which allows us to logically describe lexical relations such as synonymy and antonymy in a way that may be psychologically well-founded. It has the further effect that distributions above the word level can be defined in the same way as word distributions, so from a theoretical perspective, we do not need to define a composition operation. 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*. Of course, the extent to which we can approximate this behaviour with actual distributions is still an open question.

7 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 would argue 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 (§4.5) and sense distinctions (§5). 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.

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.

Calude, Andreea & Mark Pagel. 2010. Selfish words: what can words in competition tell us about language change? In *Proceedings of Re-thinking synonymy: semantic sameness and similarity in languages and their description*, Helsinki, Finland.

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.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge, England, UK: Cambridge University Press.

Dagan, Ido, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein & Carlo Strapparava. 2006. Direct Word Sense Matching for Lexical Substitution. In *Proceedings of COLING-ACL 2006*, Sydney, Australia.

Dilin, Liu. 2010. Construing and Conventional Usage in the Use of Synonymy: An Analysis of Two sets of (Near)-synonymous Nouns Using Both Corpus and Elicited Data. In *Re-thinking synonymy: semantic sameness and similarity in languages and their description*, Helsinki, Finland.

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*: The 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)* COLING '65, 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.
Herbelot, Aurelie & Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)* x, Oxford, England, UK.
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.

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.

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

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.

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.

Szpektor, Idan, Hristo Tanev & Ido Dagan. 2004. Scaling Web-based Acquisition of Entailment Relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP04)*, 41–48. Barcelona, Spain.

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
Institut für Linguistik
Karl-Liebknecht-Straße 24-25
D-14476 Golm, Germany
aurelie.herbelot@cantab.net