# 24

# Word Senses

KRISTER LINDÉN

*How many angels can dance on the point of a very fine needle,*
*without jostling one another?*

— Isaac D'Israeli (1766-1848)

What is the meaning of a word? Unless one believes that we are born with an innate set of meanings waiting to find their corresponding expression in language, another option is that we learn the meaning of a word by observing how it is used by the language community we are born in. Some usages find their way into dictionaries and become established word senses. In order to understand what constitutes a word sense, we can look at the criteria lexicographers use when they decide that a word usage is a word sense and record it in a dictionary for future generations.

## 24.1 Language Philosophy

From a machine learning point of view Wittgenstein's suggestion (Wittgenstein, 1953) that "*the meaning of a word is its use in the language*" sounds plausible, because there is nothing else for a machine to observe. This view of meaning was made more specific by Harris, when he proposed that words with similar syntactic usage have similar meaning (Harris, 1954, 1968).

Even if we accept that the *potential* usage of words is unlimited, we are mainly interested in *real* usage when we learn to identify similarities or differences of word meaning. The real usage is prone to fluctuations and idiosyncracies, viz. usage preferences, of different language communities. A language community is any group of individuals who communicate. Some usage preferences become recognized by most communities of a language, a process known as lexicalization. Lexicalization progresses differently in dif-

*Inquiries into Words, Constraints and Contexts*.
Antti Arppe et al. (Eds.)

ferent communities of a language giving rise to, e.g., synonyms.

The usage preferences as they manifest themselves in real usages characterize similarity or difference of word meaning. If someone says “Shoot!” when a bear is attacking, it is emotionally quite different from the same command when a small bird is flying by, although both require some weaponry. However, a reporter can shoot a question without extra equipment. For most usages of a written word, we do not have access to the full context, so there may be essential differences in other aspects than those in the text presented to a computer. Indirectly, by observing other usages of words in the context, it may still be possible for a computer to group the usages of *shoot* in ’shoot a bear’, ’shoot a bird’, and ’shoot a question’ into two main groups of shooting with and without weapons. Then we present the machine with ’shoot a bullet’ and expect the *bullet* to be more like a *question* than a *bear*, because in fact the main division does not really depend on the presumed weapon, but whether the direct object of *shoot* is animate or inanimate. We call this distinction a semantic feature. A multiple-inheritance taxonomy of such features is a feature structure. The animate and inanimate distinction is not fixed for every word, but may lend itself to modification or underspecification as in ’shooting stars’. A machine making observations based on a limited amount of samples of the real usage of a word in written text will end up with a piecewise approximation of features such as animate and inanimate.

## 24.2 Enumeration vs. Generation

The simplest way to create a dictionary of word senses is to enumerate each sense separately. If no further information is provided about how the senses are related, this representation requires each new sense to be manually added. A more flexible representation is presented by Pustejovsky (1998), a generative lexicon (GL), where the word senses are generated through the unification of feature structures guided by an inheritance system for the argument, event and qualia structures.

The GL is sometimes seen as a fundamentally different approach from the idea of dictionaries or lexicons as a simple enumeration of word senses, because the theory on generative lexicons claims that the GL also accounts for novel uses of words. Kilgarriff (2001) tested this claim on a set of corpus words and found that most of the novel or non-standard usages were unlikely to be accounted for by any GL, i.e., those usages that were not accounted for in a regular dictionary. The main benefit of a large-scale dictionary based on the GL theory would be that similar distinctions would consistently be made throughout the dictionary for all words with similar or related usages.

From a computer programming point of view, it is not particularly surprising that a lexicon program, i.e., a GL, is more flexible than a list of word

descriptions, more consistent and more compact, but equally unimaginative. In addition, as the GL grows, it is likely to be more unpredictable and more difficult to maintain. A GL comes with all the benefits and drawbacks of a large computer program and as such it covers only the words and senses it has been either intentionally or unintentionally programmed to cover.

## 24.3 The Origin of Features

A more fundamental problem related to language learning and child language acquisition is how we learn to associate meaning with sound sequences or words. We do not get closer to a solution for this problem by dividing a word into semantic features, because then we have to ask where the features come from or how they become primitives of the lexicon.

Interesting research on how meaning is associated with sound sequences has been done by Kaplan (2001) in his simulation of a robot society communicating about positions of several colored figures, i.e., circles, triangles and squares, on a white board using a Wittgensteinian language game. He was able to demonstrate that, when several stable language communities had evolved, synonymy arose. When the communities were in sporadic interaction, the communities kept their own words for the concepts but were able to understand other variants. By inspecting the robots he could determine that they had words for colors, shapes and relative positions. The robot simulations indicate that with suitable and not too complicated models, language can be learned from scratch in a language community interacting with the external world.

Research by (one of Harris' students) Gleitman (1990, 2002) on child language acquisition indicate that children learn nouns with external references before they learn verbs and then start distinguishing between different argument structures of the verbs. Her research supports the assumption that the meaning of verbs is tightly tied to their argument structure. The child language research gives some psychological relevance to the GL approach indicating that a GL is not merely a way of compressing the lexicon description.

If we accept that features and the meaning of features can be induced through language usage in a language community, a full-scale GL for some application would be an interesting effort both as a collection of linguistic knowledge and as a benchmark for future automatically induced vocabularies. It is quite likely that for some time to come high-performing computational lexicons will be partly hand-made with a generative component and a trainable preference mechanism[1]. A well-designed linguistically motivated

[1]On a parallel note, we quote Kohonen's personal comment on his self-organizing maps: "Once it has been shown that a map always organizes regardless of how random the initial state is, there is no need to show this every time. It is quite acceptable to speed things up by starting

GL with a trainable preference learning mechanism might be a good candidate for how to organize a word sense lexicon. There is no need for a computer to always learn the lexicon from scratch, despite the fact that this seems to be the way nature does it.

## 24.4 Recording Word Senses

New words and concepts arise at a steady pace and old words become associated with new meanings, especially in technology and biotechnology which are currently the focus of intense research efforts. In these areas specialized efforts like named entity recognition aim at identifying the meaning of new terms in the form of abbreviations, nouns and compound nouns by looking at their context. These entities are typically classified into names, dates, places, organizations, etc. Named entities and word senses represent two different aspects of the same problem. Named entities are usually new, previously unseen items that acquire their first word sense, whereas word sense discovery and disambiguation typically have assumed that words have at least two meanings or word senses in order to be interesting. It is, however, likely that the mechanism or process that attaches the first word sense to a string is the same as the one that later attaches additional meanings or word senses to the same string either by coincidence, i.e., homonymy, or by modifying some existing meaning, i.e., polysemy.

Other work on this theme distinguishes different word senses when a word gets different translations (Resnik and Yarowsky, 2000) so that the sense identification problem merges with finding appropriate translations. This analogy can be taken further, because finding the first word sense is in some ways equivalent to finding the first translation, which is especially important for cross-lingual information retrieval in the same areas where named entity recognition is important. A method which significantly outperforms previously known comparable methods for finding translations of named entities in a cross-lingual setting has been proposed by the author (Lindén, 2004, 2005 forthcoming).

As Kilgarriff (2003b) points out, automatically identifying a word's senses has been a goal since the early days of computational linguistics, but is not one where there has been resounding success. He suggests that the underlying problem may be unclarity as to what a word sense is (Kilgarriff, 1997). A word might not have been seen in a context because it is not acceptable there, or it might not have been seen there simply because the corpus was not big enough (Kilgarriff, 2003b). In the following, we will first look at the frequency aspect and then at the acceptability aspect.

---

from an educated guess."

### 24.4.1 Frequency Distribution

Where a lexicographer is confronted with a large quantity of corpus data for a word, then, even if all of the examples are in the same area of meaning, it becomes tempting to allocate the word more column inches and more meanings, the lexicographer Kilgarriff admits in (Kilgarriff, 2004) and considers the words *generous* and *pike* as examples:

> *Generous* is a common word with meanings ranging from generous people (who give lots of money) to generous helpings (large) to generous dispositions (inclinations to be kind and helpful). There are no sharp edges between the meanings, and they vary across a range. Given the frequency of the word, it seems appropriate to allocate more than one meaning, as do all of the range of dictionaries inspected. *Pike* is less common (190 BNC occurrences, as against 1144) but it must be assigned distinct meanings for fish and weapon (and possibly also for Northern English hill, and turnpike, depending on dictionary size), however rare any of these meanings might be, since they cannot be assimilated as minor variants. Pike-style polysemy, with unassimilable meanings, is the kind that is modeled in this paper. Where there is generous-style ambiguity, one might expect less skewed distributions, since the lexicographer will only create a distinct sense for the 'generous disposition' reading if it is fairly common; if the lexicographer encounters only one or two instances, they will not. Polysemy and frequency are entangled.

In the same article, Kilgarriff (2004) observes that the dominance of the most common sense increases with $n$, the frequency of the word. In additional corpus data, we find additional senses for words. Since a majority of the words are monosemous[2], finding additional senses for them dominates the statistic. On the average, the proportion of the dominant sense therefore increases with $n$ simply because the proportion of the first sense, $(n-1)/n$, compared to that of the additional sense, $1/n$, increases with $n$. He proceeds to demonstrate that the distribution of word senses roughly follows a Zipfian power-law similar to the well-known type/token distribution (Baayen, 2001, Zipf, 1935). Kilgarriff uses the sense-tagged SemCor database (Mihalcea, 2004) for empirical figures on the proportion of the most common sense for words at various frequencies, and compares the empirical figures with the figures his model predicts when initialized with the word frequency distribution from the British National Corpus (BNC) (Burnard, 1995). The fit between the SemCor and the predicted figures makes it believable that word frequencies and word sense frequencies have roughly similar distribu-

[2]WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet contains approximately 126,000 monosemous words with as many word senses, and 26,000 polysemous words with 78,000 word senses (Miller et al., 2003).

tions and that we can expect the skew to become more pronounced for higher values of $n$.

The conclusions we can draw from Kilgarriff (2004) are that a large-scale domain-independent word sense disambiguation system, which always chooses the most common sense out of two or more senses, will over time perform accurately in 66–77 % of the ambiguous cases based on the weighted average of the SemCor figures, or even in 66–86 % of the cases according to the figures predicted by the larger BNC corpus model. For high-frequency words, the ambition of a lexicographer to account for all the source material rather than for all the senses is a partial explanation for why some word senses are difficult to disambiguate even for humans. If such senses were disregarded, the higher predicted proportions of the dominant sense may in fact be more valid for the high-frequency words. Another implication of the Zipfian distribution is that over time all words are likely to appear in most contexts with a very low probability, and in practice most word senses will never have been seen more than once in any specific context.

### 24.4.2 Acceptability in Context

As soon as we start limiting the acceptability of words in certain contexts, we begin losing creative language use. One possibility is to relate the contents of a sentence to the world we live in, in order to estimate the plausibility of the sentence. However, this will complicate matters, because we then also have to model the plausibility of events in the world. An approximation of how objects and events of the world relate to one another is provided by an ontology. Unfortunately, there is yet no world-wide ontology around, but we have fairly large thesauri.

The difference between a thesaurus and an ontology is that the former deals with words and their relations observable in language use and the latter deals with objects and their relations in the world we live in. To high-light the distinction, we can consider the famous quote “Colorless green ideas sleep furiously” by Chomsky (1957). From a purely language use perspective this full sentence is unexpectedly likely occurring more than 5,700 times on the world-wide web. It is so common that it can be regarded as idiomatic. From an ontological perspective, the fact that it has been repeated into idiomhood by the world’s linguists does not make its content more plausible. Compositionally it still means little, but contextually it is a very pregnant construction. However, people tend to speak and write more often about things they have or would like to have experienced than they spend time producing and repeating random sequences of words, so the language we can observe is a noisy reflection of the relations between objects in the world. As a consequence, the difference is not so wide between a thesaurus constructed from observations of language use and an ontology constructed from observations of the world.

A bigger practical problem is that thesauri usually do not contain well-defined word senses that we could use for plausibility judgments. In an effort to clarify the relation between words and their multiple meanings Kilgarriff (2003a) tries to explain why thesauri do not really contain word senses. The first priority of authors of thesauri is to give coherent meaning-clusters, which results in quite different analyses from those in dictionaries, where the first priority is to give a coherent analysis of a word in its different senses (Kilgarriff and Yallop, 2000). From a practical point of view, if we wish to use a thesaurus for a natural language processing (NLP) task, then, if we view the thesaurus as a classification of word senses, we have introduced a large measure of hard-to-resolve ambiguity to our task (Kilgarriff, 2003a). For this reason Kilgarriff claims that, even though Roget may have considered his thesaurus (Roget, 1987) a simple taxonomy of senses, it is better viewed as a multiple-inheritance taxonomy of words.

The direct consequence of Kilgarriff's argument is that a thesaurus is perhaps useful as a backbone for a generative lexicon, but as such the words in a thesaurus are ambiguous. Kilgarriff's argument is easier to understand if we keep in mind that the meaning of a word is defined by the contexts in which it occurs. The real problem is that a meaning-cluster in a thesaurus seldom includes the common contexts in which the words of the meaning-cluster occur. So what can we use a thesaurus for? Systems which try to discover word senses, also classify words based on their context into maximally coherent meaning-clusters, i.e., thesauri can serve as test beds for automatic word sense discovery systems. The somber consequence of Kilgarriff's argument is that for NLP systems the words in a meaning-cluster are in fact an epiphenomenon[3]. The valuable part is the context description by which the words were grouped. The context description is a compact definition of the meaning of the word cluster and this is the part that is usually made explicit in a regular dictionary analyzing the senses of a word. It is the context description that can be used for determining the acceptability of the word sense in various contexts.

## 24.5 Word Sense Dictionary Specification

If we use a generative lexicon to determine the acceptability of a word sense in context and the lexicon provides hard constraints, we will end up not covering creative language use after all. We could, however, account for creative lan-

[3]This is not to say that word sense and thesaurus discovery efforts are futile. Word lists are primarily intended for consumption by systems that are capable of filling in the appropriate context descriptions themselves, e.g., human beings. A central issue in information retrieval (IR) research is to devise strategies which cope with missing context. This may partially explain why IR often seems to have more to offer thesaurus makers than the other way around, see (Sanderson, 2000).

guage use by basing plausibility judgments[4] on observable language. Ideally, a lexicon provides structure and soft constraints based on context descriptions giving more plausibility to more likely objects and events.

To summarize the discussion of the previous sections, we can set up a general wish list of what a context description of a word sense in an ideal lexicon should contain, loosely based on the idea of a generative lexicon (Pustejovsky, 1998): *part of speech* categories, *argument structure* of arguments and adjuncts, *event structure* for the argument structure, *qualia structure* describing an object, its parts, the purpose and the origin of the object, *interlexical relations*, e.g., synonymy, antonymy, hyponymy, entailment, translation, *plausibility estimate* by providing all of the above with frequency or probability information[5].

An example of the plausibility information the lexical model needs to incorporate is given by Lapata and Brew (2004), where they highlight the importance of a good prior for lexical semantic tagging. They find a prior distribution for verb classes based on Levin (1993), and they obtain their priors directly from subcategorization evidence in a parsed but semantically untagged corpus.

Another example is the prevalence ranking for word senses according to domain, which should be included in the generative lexical look-up procedure. The sense distributions of many words depend on the domain. Giving low probability to senses that are rare in a specific domain permits a generic resource such as WordNet to be tailored to the domain. McCarthy et al. (2004) present a method which calculates such prior distributions over word senses from parsed but semantically untagged corpora.

## 24.6 Conclusion

In text we can observe word forms which through morphological analysis get a base form. A base form may have several meanings which together form a lexeme. An explicit *meaning–base form* pair, i.e., a word sense, is an artifact we cannot observe directly. We can only observe word usages. The only evidence we have for a word sense is found in a dictionary via the definitions and glosses provided by a lexicographer reflecting meaningful groups of word usages.

[4] A plausibility judgment is at least a weak partial ordering of the relative plausibility of statements.

[5] From a Bayesian statistics point of view we would have prior linguistic information combined with the posterior information provided by corpus data. Before we have seen any data, our prior opinions about what the true relationships might be can be expressed in a probability distribution over the feature structure weights that define the relationships. After we look at the corpus data (or after our lexicon is adapted to the data), our revised opinions are captured by a posterior distribution over the feature structure weights.

We have briefly described the criteria lexicographers use when they decide which word usages constitute a word sense. The fact that the bulk of all language use is a reflection of the world we live in, makes some word senses of a word dominant. Most previously unseen word usages are creative simply because they are unexpected or surprising at the time. A natural language processing (NLP) system needs to recognize that a usage is unexpected. However, the context in which the usage appears is what the word means and should be recorded for future reference, e.g., telephones used to be stationary until the advent of mobile phones, so a sentence like "He walked down the street talking on the phone" was implausible 30 years ago, but is now highly likely and the walking-talking context has become part of the meaning of a telephone.

We have argued that word meaning is not discrete. However, the meaning of words is quantized into word senses in a dictionary. If we need a common world view, we can refer to a sense inventory of an agreed upon dictionary, otherwise we can as well compare word contexts directly.

## References

Baayen, Harald R. 2001. *Word Frequency Distributions*, vol. 18 of *Text, Speech and Language Technology*. Dordrecht: Kluwer Academic Publishers.

Burnard, Lou. 1995. *Users' Reference Guide for the British National Corpus, version 1.0*. Oxford University Computing Services, Oxford, UK.

Chomsky, Noam. 1957. *Syntactic structures*. Janua Linguarum Series Minor, Volume 4. The Hague, The Netherlands: Mouton de Gruyter.

Gleitman, Lila R. 1990. The structural sources of verb meaning. *Language Acquisition* 1:3–55.

Gleitman, Lila R. 2002. Verbs of a feather flock together II: The child's discovery of words and their meanings. In B. E. Nevin and S. B. Johnson, eds., *The Legacy of Zellig Harris: Language and information into the 21st century*, vol. 1: Philosophy of science, syntax and semantics of *Current Issues in Linguistic Theory*, pages 209–229. John Benjamins Publishing Company.

Harris, Zellig S. 1954. Distributional structure. *Word* 10:146–162.

Harris, Zellig S. 1968. Mathematical structures of language. *Interscience Tracts in Pure and Applied Mathematics* 21(ix):230 pp.

Kaplan, Frédéric. 2001. *La Naissance d'une Langue chez les Robots*. Collection Technologies et Culture. Paris, France: Hermes Science Publications.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31(2):91–113.

Kilgarriff, Adam. 2001. Generative lexicon meets corpus data: the case of non-standard word uses. In P. Bouillon and F. Busa, eds., *The Language of Word Meaning*, pages 312–328. Cambridge: Cambridge University Press.

Kilgarriff, Adam. 2003a. Thesauruses for natural language processing. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering*. Beijing: Beijing Media Center.

Kilgarriff, Adam. 2003b. What computers can and cannot do for lexicography, or Us precision, them recall. Tech. Rep. ITRI-03-16, Information Technology Research Institute, University of Brighton. Also published in Proceedings of ASIALEX.

Kilgarriff, Adam. 2004. How dominant is the commonest sense of a word? In P. Sojka, I. Kopeček, and K. Pala, eds., *Proceedings of TSD 2004, Text, Speech and Dialogue 7th International Conference*, vol. 2448 of *LNAI*, pages 1–9. Brno, Czech Republic: Springer-Verlag, Berlin.

Kilgarriff, Adam and Colin Yallop. 2000. What's in a thesaurus? In *Proceedings of LREC 2000, the 2nd International Conference on Language Resources and Evaluation*, pages 1371–1379. Athens.

Lapata, Mirella and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics* 30(1):45–75.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: University of Chicago Press.

Lindén, Krister. 2004. Finding cross-lingual spelling variants. In *Proceedings of SPIRE 2004, the 11th Symposium on String Processing and Information Retrieval*. Padua, Italy.

Lindén, Krister. 2005 forthcoming. Multi-lingual modeling of cross-lingual spelling variants. *Information Retrieval* .

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Automatic identification of infrequent word senses. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1220–1226. Geneva, Switzerland.

Mihalcea, Rada. 2004. Software and data sets – semcor. [http://www.cs.unt.edu/ rada/downloads.html#semcor].

Miller, George A., Christiane Fellbaum, Randee Tengi, Susanne Wolff, Pamela Wakefield, Helen Langone, and Benjamin Haskell. 2003. Wordnet – a lexical database for the English language. [http://www.cogsci.princeton.edu/∼wn/index.shtml].

Pustejovsky, James. 1998. *The Generative Lexicon*. The MIT Press.

Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering* 5(3):113–133.

Roget, Peter Mark. 1987. *Roget's Thesaurus*. Longman, Longman Edition edit by Betty Kirkpatrick edn. Original edition 1852.

Sanderson, Mark. 2000. Retrieving with good sense. *Information Retrieval* 2(1):49–69.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell. Translated by G. E. M. Anscombe.

Zipf, George Kingsley. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Boston, USA: Houghton Mifflin.