Selectional Preferences, Corpora and Ontologies

Michal Měchura

M.Phil. in Speech and Language Processing

Dissertation

September 2008

Trinity College, University of Dublin

Wordcount: 14,394

Declaration

I declare that this dissertation has not been submitted as an exercise for a degree at this or any other university and that it is entirely my own work. I agree that the Library may lend or copy this dissertation on request.

Signed:

Date:

Acknowledgement

I wish to thank my supervisor, Elaine Uí Dhonnchadha, for her gentle guidance throughout. I am also grateful to Kevin Scannell of Saint Louis University and to Carl Vogel of Trinity College, University of Dublin who have provided comments on an earlier draft. Last but not least, a word of thanks to my significant other Ivanka for her endless patience while I was not thinking or talking about much else than this dissertation.

Abstract

Selectional Preferences, Corpora and Ontologies

Michal Měchura

This work presents a technique for exploring the selectional preferences of words in a semi-automatic way. The technique combines corpora with ontologies such as WordNet.

The term *selectional preference* denotes a word's tendency to co-occur with words that belong to certain lexical sets. For example, the adjective *delicious* prefers to modify nouns that denote food and the verb *marry* prefers subjects and objects that denote humans. This work develops techniques for associating corpus-attested selectional preferences with concepts in an ontology. It shows how lexical sets can be derived from ontologies and how corpus-extracted collocates of a word can then be aligned with these lexical sets to reveal any selectional preferences the word has.

An additional contribution provided here is an insight into the limitations of this method. The work presents evidence for the conclusion that aligning selectional preferences with an ontology is useful for some purposes, but fundamentally inaccurate because currently existing ontologies do not accurately reflect the mental categories evoked in selectional preferences.

Contents

Chapter 1

Introduction

1.1 A brief history of selectional preferences

Selectional preferences were an important subject of interest in the early development of generative grammar. Noam Chomsky incorporated them into his syntactic theory where they appeared in context-free rules as features, such as [+ANIMATE], [-ABSTRACT], to constrain the choice of lexical items (Chomsky, 1965, pp. 113ff and 153ff). To this day, an account of the phenomenon of selectional preference is part of many grammatical frameworks such as HPSG (Soehn, 2005).

Some aspects of selectional preferences have been debated extensively in the literature, for example the question whether selectional preferences are part of syntax, semantics or pragmatics (Katz and Postal, 1964; Chomsky, 1965), whether they are predictable from the meanings of words (McCawley, 1976) and what their status is with respect to a logical theory of language (Seuren, 1985). However, not much attention has been given to the descriptive cataloguing of the selectional preferences of individual words. As the manual discovery of selectional preferences is labour-intensive and slow, it was not long until the idea emerged (Resnik, 1993) that selectional preferences could be induced automatically from corpus data in combination with a pre-existing ontology such as WordNet (Fellbaum, 1998). Resnik and his followers have created algorithms that estimate which categories in

WordNet best correspond to a word's selectional preferences—for an overview see Light and Greiff (2002). This line of research is continued here.

In all work which attempts to align selectional preferences with an ontology, there has been an unspoken assumption that it will always be possible to find a place in the ontology for every selectional preference. Upon closer scrutiny, this does not in fact appear to be so straightforward, as will be shown later. So, in addition to developing practical tools, an objective of this work is to obtain insights of theoretical importance into the nature of the relationship between selectional preferences and ontologies.

1.2 Selectional preferences or selectional restrictions?

In the 1960s and the 1970s, the term used in the literature to refer to the phenomenon was *selectional restrictions*. In modern corpus-based work, the preferred term seems to be *selectional preferences*—perhaps a reflection of the fact that when examples of real-world language in use are examined, the "restrictions" appear rather indeterminate and not as restrictive as the term would imply. In this work I have decided to follow the contemporary practice and use the term *selectional preferences*, but with the qualification that the truth is somewhere in between. What appears as indeterminacy can often be explained by phenomena such as metonymic extension and metaphorical usage. The apparent difficulty in "pinning down" selectional preferences will be dealt with in a later chapter.

1.3 Map of the dissertation

This work attempts to achieve two interrelated objectives, a theoretical one and a practical one. The theoretical objective is to investigate the relationships between selectional preferences on the one hand and semantic ontologies on the other. The practical objective is to build a software tool, called SenseMaker, which allows linguists to explore the relationships. The remainder of the dissertation proceeds

according to the following scheme.

Chapter 2 is mainly theoretical. It provides an overview of the nature of selectional preferences and of their significance in natural-language processing and in lexicography. The material in this chapter draws on previously published literature as well as on my own observations.

Chapter 3 will attempt to answer whether (and to what extent) selectional preferences of individual words correspond to the categories contained in ontologies such as WordNet. This prepares the ground for the next chapter by suggesting reasons why it may be useful to find such correspondences.

Chapter 4 is more technical and documents the design principles which were involved in the building of SenseMaker, a tool for discovering selectional properties of words. In particular, it deals in some detail with the way I have derived lexical sets from the various semantic relations in WordNet.

Chapter 5 presents several case studies of words whose selectional properties I have investigated using SenseMaker.

Chapter 6 looks at the results of the case studies and makes some remarks with respect to the suitability of lexical ontologies as means for capturing and explaining selectional preferences.

Chapter 7 makes suggestions for further work, both in terms of additional research that needs to be done to understand the nature of selectional preferences, and possible applications in natural-language processing and in lexicography that result from this work.

In the interest of clarity, the following is a brief listing of areas which the present work does *not* deal with. First of all, this project does not work with semantically annotated corpora. Any semantic data is supplied by an external ontology such as WordNet. Neither is this a corpus annotation project. Semantically annotated corpora will noy be an outcome of the project. Last but not least, the project

does not aim to cluster words semantically purely on the basis of their distributional properties.

1.4 Notational conventions

This work uses the following conventions of notation. When word forms are mentioned regardless of their sense, they appear in italic type, such as *build* and *construct*. The concepts designated by words are mentioned in small capitals, such as BUILD. The term "concept" as used here roughly corresponds to a synset in WordNet. The details of WordNet, including an explanation of what is a synset, are presented in a later chapter.

When quoting examples, the hash character # is prefixed to those which are unacceptable on account of a violation of a selectional preference, and an asterisk * is prefixed to those which are ungrammatical.

All Internet URLs quoted were last verified as accessible on 15 September 2008.

Chapter 2

A review of selectional preferences

In this chapter, the phenomenon of selectional preference is reviewed, drawing on relevant literature as well as the author's own observations. The goal is to build up a solid understanding of the phenomenon before the dissertation moves on to its main purpose.

2.1 Selectional preferences are arbitrary

As explained, the selectional preferences of a word are the word's tendency to co-occur with words of certain semantic classes in a given role, such as the tendency of *eat* to take words denoting FOOD as its direct objects. In some cases, the selectional preference of a word can be explained away as a logical consequence of its meaning (McCawley, 1976): the meaning of *eat* appears to imply that its direct objects must be edible entities, that is, FOOD. In other cases, however, a word's selectional preferences seem to be less logical and more arbitrary: we can *join* the army as well as *enlist* in it, but we can only *join* a political party – to say we *enlist* in one is odd. There is no obvious reason why this has to be so, it merely is so. In this work, all selectional preferences are treated as essentially arbitrary, even in cases when

their motivation may in principle be found in the meanings of the words involved.

2.2 Selectional preferences are language-specific

Looking at selectional preferences cross-linguistically, one finds that they are often language-specific. It may seem that pairs of translation equivalents must share the same selectional preferences, as indeed they often do: the kinds of things we say we *eat* in English are more or less the same kinds of things we say we *essen* in German, *itheann* in Irish and *jíme* in Czech. However, there are pairs which do differ in their selectional preferences. The German translation equivalent of *drive* (as in *drive a vehicle*) is *fahren* but it has a wider scope than *drive* as it can combine with motorcycles and bicycles as well. The German equivalent of *ride* is *reiten* but it has a narrower scope than *ride* as it cannot combine with motorcycles or bicycles. The examples in (1) and (2) demonstrate this.

(1) a. (i) Kim drives a truck/ car.

(ii) #Kim rides a truck/ car.

b. (i) #Kim drives a motorcycle/ bicycle/ horse.

(ii) Kim rides a motorcycle/ bicycle/ horse.

(2) a. (i) Ute fährt ein(en) Lastwagen/ Auto/ Motorrad/ Fahrrad.
Ute drives a truck/ car/ motorcycle/ bicycle.

(ii) #Ute fährt ein Pferd.
Ute drives a horse.

b. (i) #Ute reitet ein(en) Lastwagen/ Auto/ Motorrad/ Fahrrad.
Ute rides a truck/ car/ motorcycle/ bicycle.

(ii) Ute reitet ein Pferd.
Ute rides a horse.

In English, *drive* has a selectional preference for motorized vehicles with more than two wheels, and *ride* has a preference for vehicles on which the rider is posi-

tioned astride, regardless of whether motorized or not, and inclusive of animals.[1] The picture is different in German: *fahren* has a preference for all vehicles, motorized or not but exclusive of animals, while *reiten* selects for animals only (Soehn, 2005). In this work, all selectional preferences are treated as essentially language-specific, even in cases when a cross-linguistic correlation might in principle hint at a language universal.

2.3 Multiple selectional preferences per word

Words often have more than one selectional preference. While *eat* has only one (FOOD), the verb *follow* has several, including PATH (*path, route, trail, track*), INSTRUCTION (*instructions, guidelines, recommendation*) and EVENT (*publication, resignation, arrest*).[2] The selectional preferences sometimes help to distinguish between what lexicographers recognize as separate senses of the word: *follow* has the sense 'happen after' when its object is an EVENT and 'move along' when it is a PATH.

While selectional preferences sometimes help to disambiguate between senses, they do not always. Consider the verb *shoot*. This has two core senses, 'kill with gun' when the object is ANIMATE (as in *the police shot the criminal*) and 'take a photograph' when it is a DEPICTION (as in *the photographer shot the picture*) (Soehn, 2005). However, there are attested examples of *shoot* + ANIMATE when clearly the photography sense is intended, as in the sentence *How do I, as a young photographer, shoot an unexperienced model?* (actual question asked at the Ya-

[1] There is an area of overlap between things that can be *driven* and things that can be *ridden*. To say that one *drives a motorcycle* is not completely unacceptable, but it is markedly less common than saying one *rides* it. There are only two occurrences of *drive* + *motorbike* (and none of *drive* + *motorcycle*) in the British National Corpus (`http://www.natcorp.ox.ac.uk/`). Regardless of whether the boundary between drivable and ridable objects is fuzzy or not, the point is that the boundary is elsewhere in English and in German.

[2] The claim that *eat* has only one selectional preference is arguable, depending on whether one wishes to count metaphorical uses such as *what's eating you?* separately or not.

hoo! Answers website). Similarly, the sense of *I love shooting blondes* is different depending on whether it is uttered by a photographer or a serial killer. This means that not only is the word *shoot* ambiguous, but the pattern *shoot* + ANIMATE is ambiguous too and only the wider co-text, along with the full gamut of a human's deductive abilities, can disambiguate it successfully. The conclusion is that while selectional preferences contribute greatly towards word-sense disambiguation, they are not the complete answer.

2.4 Selectional preferences are specific to syntactic roles

In a sentence, constituents stand in different roles to each other. For example, in *your mother is concerned about your health*, the two noun phrases *your mother* and *your health* stand in two different roles to the predicate *be concerned*: *your mother* is its subject and *your health* is its (indirect) object, embedded inside a prepositional phrase headed by *about*. The predicate has different selectional preferences for each role. The subject can be HUMAN, ORGANIZATION and even PUBLICATION (as in *the book is concerned with the conflict between science and religion*). The indirect object, when headed by *about*, can be a STATE (*concerned about your health*), a CHANGE TOWARDS SOMETHING BAD (*concerned about the loss of confidence*), and so on. If the indirect object is headed by a different preposition, say *with*, then the selectional preferences are a little different, including SUBJECT MATTER (*the book is concerned with the conflict between science and religion*). In this text, when we talk about a word having a selectional preference, we mean that it has the selectional preference for a particular role. Formally, a selectional preference is a three-tuple consisting of the predicate, the role, and the semantic category, such as <*be concerned*, subject, HUMAN> (Light and Greiff, 2002). This three-tuple notation is used in this work occasionally when absolute clarity is required.

2.5 Do all words have selectional preferences?

A typical treatment of selectional preferences in the literature is to treat them as properties of predicates (mostly verbs and adjectives). Under this view, predicates impose certain selectional preferences on their arguments, effectively answering questions such as: What kind of thing can *eat*? What kind of thing can be *eaten*? What kind of thing can be called *desperate*? But it is possible to look at selectional preferences in the other direction as well, as arguments selecting for predicates (Light and Greiff, 2002, p. 270): we can say about a sandwich that it is *delicious* or *wilted* but not that it is *desperate*. In fact, the substance of selectional preference is *semantic compatibility* between predicates and their arguments.

How specific are selectional preferences? Some words are very particular indeed as to what they will combine with, as in the case of *drive* which requires that its direct objects be motorized vehicles of a particular kind: cars, buses and trucks but not motorcycles or bicycles. Then there are predicates which have a fairly wide range: to *sleep* you must be an animate being,[3] but ANIMATE is a fairly high-level category. And then there are predicates which are happy to combine with arguments of any kind at all, for example there seems to be no end to the kinds of things you can describe as *important*. Resnik (1996) has observed that the specificity of the preference predicts whether a transitive verb can be used intransitively: you can say *John ate* without saying what he ate because *eat* selects for FOOD which is a fairly specific category. But you cannot say *John made* because *make* selects for pretty much anything, its selectional preference is too general and so its filler must not be left unspecified.

2.6 Selectional preferences are extendable

So far, it would appear that the selectional preference of a word can be formulated precisely by way of a definition, as has been done above in the case of *drive* and

[3] Leaving metaphorical extension such as machines and computers aside for the moment.

ride (e.g. "motorized vehicles with more than two wheels"). However, the situation is complicated by the interference of semantic relations such as metonymy. The verb *drink* has a selectional preference not only for LIQUID (*she likes drinking foreign beer*) but also for CONTAINER OF LIQUID (*she drank a full bottle of beer before she left the house*) and QUANTITY OF LIQUID (*if you drink more than five litres of beer a day*).

This phenomenon has been studied extensively in the Generative Lexicon tradition (Pustejovsky, 1995) under the terms *type shifting* and *type coercion* (Pustejovsky, 1993). The general idea is that a constituent (word, phrase, ...) has a default denotation as well as a set of ways in which the denotation can be "shifted" by context. We will not go beyond stating that this phenomenon exists and that it is responsible for the extendibility of selectional preferences.

2.7 Selectional preferences are context-sensitive

Additionally, selectional preferences are context-sensitive in the sense that almost any imaginable item can fit the preference provided it is construed appropriately in the co-text. Consider the sentences in (3).

(3) a. # Pete ate his motorcycle right away.

b. The twins were given chocolate motorcycles for winning the race. Pete ate his motorcycle right away but Marty said he wanted to keep his for later.

In isolation, the sentence in (3-a) is odd because it violates *eat*'s selectional preference for edible entities. However, this can be turned into a perfectly acceptable sentence if the referent of *motorcycle* is established as something edible by the preceding text, as it is in (3-b).

The lesson to learn from these examples is that the fulfilment or otherwise of a selectional preference by a word may not always be determinable from its usual

reference: *motorcycle* usually refers to VEHICLE but it can be shifted by context to anything at all, including FOOD. The precise referent of the word must sometimes be deduced from the co-text (*chocolate motorcycles* are edible, therefore they are FOOD) or by following lines of conventional metonymic extension (LIQUID → CONTAINER). It is the referent that fulfils the selectional preference, not the word. Once again, only the full gamut of a human's deductive abilities is the final arbiter of whether the selectional preference is fulfilled or not.

2.8 Selectional preferences and metaphor

A sentence may contain an apparent violation of a selectional preference and still be interpretable when the violation simply forces a metaphorical reading, as in example (4) where *drink*'s preference for animate subjects is violated in the literal reading but fulfilled in the metaphorical one.

(4) My car drinks fuel like mad.

Some metaphorical exploitations of a selectional preference may become conventionalized over time and become a selectional preference in their own right, as has presumably happened to the verb *devour* in (5) (Hanks, 2008).

(5) She devoured the book in just a couple of hours.

The metaphorical usage may even outweigh the original non-metaphorical usage in terms of frequency of use. The combination <*devour*, object, BOOK> is the most frequent in the British National Corpus [4], outranking the likes of <*devour*, object, MEAL> and <*devour*, object, PREY>.

[4] http://www.natcorp.ox.ac.uk/

2.9 Applications of selectional preferences in natural-language processing

Knowledge of selectional preferences can help resolve ambiguities in several language-engineering tasks. In parsing, attachment ambiguities sometimes arise, as in the sentence in (6).

(6) She saw the man with the hat.

Here, the prepositional phrase *with the hat* is to be attached as a modifier to *the man* because attaching it to *saw* would violate the verb's selectional preference for OPTICAL INSTRUMENT.

There are applications in speech recognition where selectional preferences can provide guidance as to how felicitous a sequence of words is. In the examples in (7), having already recognized *they ate*, the system can conclude that *peaches* is more likely than *beaches* because it satisfies the verb's selectional preference for FOOD (Light and Greiff, 2002, p. 272).

(7) a. They ate peaches.
b. They ate beaches.

As a last example, selectional preferences are useful for word-sense disambiguation. If we wish to disambiguate the word *meat* in the sentence (8) between its sense of 'flesh' and its sense of 'gist' (as in *the meat of the argument*), then the verb *eat*'s selectional preference for FOOD will support the 'flesh' reading (Light and Greiff, 2002, p. 272).

(8) Vegetarians don't eat meat.

2.10 Applications of selectional preferences in lexicography

The significance of selectional preferences is well known in natural language processing. In this work, however, we concentrate on lexicography. Selectional preferences play an important part in the writing of dictionary entries, both monolingual and bilingual.

It has become a matter of policy in recent years for some dictionaries to adopt definition-writing styles which make specific reference to selectional preferences. For example, the definition of one of the senses of *sustain* in the Macmillan English Dictionary for Advanced Learners (Rundell, 2007) is 'to experience loss, injury, damage etc'. It exemplifies the typical objects of the verb.

Selectional preferences also play a part in pointing out the differences between words that appear to be synonyms. The English verbs *adore* and *worship* appear to be synonyms, meaning roughly 'love and respect very much'. An analysis of their selectional preferences, however, reveals differences in terms of the types of things that usually appear as their objects. While *God* and other kinds of DEITY are the most common objects of *worship*, the verb *adore* doesn't display quite the same preference for DEITY. In fact, it appears to display a dispreference for it. This suggests that if one wishes to express the idea that he or she 'loves and respects God very much', then *worship* is the verb to choose, rather than its apparent synonym *adore*. This is a fact a lexicographer needs to be aware of and needs to consider for inclusion in the dictionary—especially if it is a production-oriented[5] dictionary for second-language learners who wish to use the target language according to convention.

While knowledge of selectional preferences helps to explain the differences between apparent synonyms, it can also help the lexicographer find new synonyms

[5] A production-oriented dictionary is one whose task is to help its user speak and write in the foreign language – as opposed to a comprehension-oriented dictionary whose task is merely to help the user comprehend the foreign language.

which may not be immediately obvious. The verbs *sustain* and *suffer* are synonyms when talking about injuries and losses of military personnel. This is a very narrow area of overlap and many dictionaries do not state it explicitly. It is, nonetheless, a fact which can be discovered by analysis and comparison of the two words' selectional preferences.

In a bilingual situation, an explicit mention of selectional preferences in the dictionary article can help the user choose the right translation equivalent, as in figure 2.1 from an English–German dictionary for English-speaking learners of German where selectional preferences are exemplified in square brackets.

sustain vt

a [*load, weight*] aushalten, tragen, [*life*] erhalten, [*family*] unterhalten, [*charity*] unterstützen, [*body*] bei Kräften halten

b [*pretence, argument, theory, effort, veto, interest, support*] aufrechterhalten, [*growth, position*] beibehalten, [*note*] aushalten, [*accent, characterization*] durchhalten, [*objection*] stattgeben

c [*injury, damage, loss*] erleiden

Figure 2.1: The dictionary entry for *sustain*, from Collins German Dictionary (Beattie *et al.*, 2004).

Selectional preferences can also help explain the differences between what may otherwise seem like straightforward translation equivalents. An example is the Irish verb *ól*, usually translated as *drink*. That is indeed its most frequent translation equivalent. However, one peculiar property of *ól* is that it can take as direct objects things like *píopa* 'pipe' and *tobac* 'tobacco', which *drink* cannot. This means that *drink* and *ól* are not straightforward translations of each other. In fact, *smoke* rather than *drink* is the only suitable equivalent of *ól* if pipes and tobacco are the direct objects. A similar example is the English verb *subscribe* and its typical Czech translation equivalent *předplatit*. In English, one can subscribe to

many things, including PERIODICAL, SERVICE, MEMBERSHIP and IDEOLOGY. In Czech, the verb *předplatit* is capable of taking all of these, except IDEOLOGY. To subscribe to an ideology, no obvious equivalent is available in Czech and a super-ordinate verb must be used, such as *souhlasit* 'agree'. These are but two examples of the wide-spread phenomenon of non-equivalence, known all too well to professional translators and other kinds of language professionals who have learnt from experience that they cannot rely on dictionaries exclusively. A detailed knowledge of the selectional preferences of the words involved, in both languages, can help the lexicographer identify situations of non-equivalence and suggest alternatives.

2.11 Finding selectional preferences

In order to write reliable dictionary entries which take proper account of selectional preferences, one must first know what those selectional preferences are. While it is possible to discover selectional preferences purely by introspection, a more objective method is to allow for one's intuition to be guided by corpus data. Currently, lexicographers usually investigate selectional preferences by consulting corpus-extracted word lists, such as the list below which constitutes the 25 most salient[6] direct objects of the verb *sustain* (extracted by the Sketch Engine[7] (Kilgarriff *et al.*, 2004) from the British National Corpus).

> injury, damage, loss, momentum, growth, wound, pedal, collusion, morale, level, life, fracture, interest, casualty, argument, rib, improvement, recovery, relationship, motivation, dialogue, commitment, mood, expansion, conviction

Having consulted the list, the lexicographer will probably notice that it can be subdivided into at least two clusters, one centring on the negatively charged ideas

[6] In the Sketch Engine, roughly speaking, salience is a measure of how frequently word A occurs as a collocate of word B, with respect to the frequency word A occurs on its own. In other words, salience is a more realistic measure of collocation than frequency.

[7] http://www.sketchengine.co.uk/

of LOSS and INJURY and the other centring on fairly positive processes such as GROWTH and LIFE. Depending on the purpose of his or her work, the lexicographer may then decide that the latter cluster needs further division into several more fine-grained subclusters such as kinds of IMPROVEMENT (*growth, recovery*) and kinds of CONTINUATION (*momentum, growth*).

Traditionally, this analysis is performed by the lexicographer manually, as is the established practice in lexicography today. In contrast, the present work attempts to answer the question, to what extent is it feasible to create a software tool that performs this clustering automatically? Ideally, one would have access to a software tool that takes as input a list of words – such as a list of the most frequently occurring direct objects of some verb – and subdivides the list into semantic clusters.

2.12 Extensional and intensional definition of selectional preferences

Effectively, selectional preferences bring about the existence of lexical sets. For example, all the words that can appear as the direct objects of *eat* constitute a lexical set. The lexical sets created by selectional preferences appear to be semantically coherent in the sense that all the members of the set are semantically similar to one another in some way: they meet some semantic criteria which qualifies them for membership in the set. This suggests that a collocational lexical set can be defined not only extensionally by giving an explicit listing of its members but also intensionally as a concept, as a bundle of semantic criteria that serves as a membership function for the set.

Ontologies such as WordNet[8] (Fellbaum, 1998) also bring about the existence of lexical sets, defined intensionally by way of definitions (such as kinds of ANIMAL, parts of HOUSE) and in the case of WordNet also extensionally by listing all

[8] http://wordnet.princeton.edu/

their members (words and multi-word items). The next obvious step is to ask, how do selectional preferences correlate with existing ontologies such as WordNet? In the ideal case, a selectional preference displayed by a word would be coextensive with a lexical set derived from WordNet: for example the set of all members and hyponyms of the synset FOOD would map directly onto the selectional preference the verb *eat* has for its direct objects. However, as we will see later, the correspondence between corpus-attested selectional preferences on the one hand and lexical ontologies on the other is almost never this straightforward.

Chapter 3

Selectional preferences in corpora

In this chapter, I attempt to answer the question whether selectional preferences can be induced from corpora, and if so, how. I will suggest that the most promising method is one which involves the crosschecking of corpus evidence against pre-existing lexical ontologies such as WordNet, as Resnik (1993) and his followers have done – for an overview see Light and Greiff (2002).

Combining a corpus with an ontology essentially brings into a corpus something which was not in it before. Ontologies are typically hand-crafted, a lot of intuition and thinking goes into deciding how categories in the ontology should be organized. Most importantly, the assignment of words into meaning groups is decided by humans and these decisions are not supported by any statistical analysis of the words' distributional properties. In other words, lexical ontologies are more a result of reasoning, less a result of observing authentic language usage.

There is, however, an element of circularity in the approach of this dissertation, or perhaps "spirality": an existing lexical resource (WordNet) is used to analyze corpus data to help us produce even better lexical resources.

3.1 Outline of the approach

The technique studied here essentially comprises three steps. Let's presume we wish to investigate what selectional preferences a verb, such as *intercept*, has on some category of its collocates, say its direct objects. First, we need to extract from a corpus a list of those collocates, such as {*call, letter, message, vessel*}. Second, we need to have access to a source of semantic data. Third, we need to cross-check the list of collocates against the semantic data source and output any generalizations that can be induced from it, such as the observation that three of the four words are kinds of COMMUNICATION.

3.2 Obtaining the corpus data

Obtaining a list of the collocates of a given word is a trivial task in most corpus-query systems such as WordSmith[1] as that is one of their primary functions. It is usually very easy to extract a list of collocates in a given position to the left or right of the pivot word, in lemmatized form and/or with part of speech tags if the corpus is tagged, along with a frequency count. I will refer to this approach as *positional* as it extracts collocates based purely on their linear position to left or right and ignores the syntax. In the positional approach, there is no guarantee that the collocates extracted will all be of the same word class, nor is there a guarantee that they will all stand in the same syntactical role with respect to the pivot word. For example, if we wish to obtain a list of the direct objects of the verb *send*, then a list of all words in the R2 position (second word to the right) will only approximate that. The list will indeed include many direct objects, such as *letter* and *package*, but will also include noise in the form of function words, adjectives, intervening adverbs and so on. In the concordance extract in figure 3.1, the italic words are words in the R2 position, and the underlined words are the words we really want to get if we are interested in the objects of *send*.

[1] http://www.lexically.net/wordsmith/

sex or dirty needles **send**	HIV *into* the blood.
signed after you have **sent**	a *payment* by cheque
For details please **send**	a *stamped* addressed envelope
In response, AI **sent**	out *information* packs
coming in, AI will be **sending**	out *information* on
Salvador and Iran but **sent**	a *clear* signal that
treatment. AI also **sent**	urgent *appeals* to the
returned, could you please **send**	them *to* Dr Que's brother
Spain. A Christmas card **sent**	to *him* by an AI member
imprisoned merely for **sending**	his *paintings* to North
because of reports he **sent**	to *the* British Broadcasting

Figure 3.1: Concordance lines for the verb *send*.

Position is only an approximate indicator of syntactic role. If a high accuracy is to be achieved in the induction of selectional preferences, the positional approach is not the ideal supplier of corpus data. To make sure we are getting at the direct-object nouns without any noise, we need a syntactically parsed corpus such as a treebank. In a treebank, it is usually possible to extract the noun phrases that constitute the direct objects of that verb, and further to extract the head nouns from those. Those head nouns can then be cross-checked against our source of semantic data to induce any common semantic features.

The problem is that treebanks are not nearly as widely available as syntactically unparsed corpora, and when they are available, they tend to be significantly smaller. Fortunately, the problem can be solved with a form of shallow parsing performed on demand. One corpus-query system that performs shallow parsing is the Sketch Engine (Kilgarriff *et al.*, 2004). The Sketch Engine searches a lemmatised, part-of-speech tagged corpus for occurrences of a given word, say some verb. Then it makes use of patterns written in CQL (Corpus Query Language)[2] to locate

[2] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

the verb's subjects, objects, and other roles, and finally compiles a list of those along with frequency counts and salience data. This method is ideal for extracting syntax-aware collocate lists with an accuracy high enough for the purposes of the probabilistic analysis performed here.

In the remainder of this work, the techniques will be exemplified on collocate lists extracted from the British National Corpus using the Sketch Engine. It should be remembered, however, that the method is equally applicable to data extracted from treebanks and, with some reservations, from "raw", unparsed corpora.

3.3 A short overview of WordNet

Having obtained a list of collocates, we want to cross-check it against an ontology. But what exactly is an ontology? The term *ontology* is used in several disciplines to mean slightly different things. In philosophy, it simply means "all that exists". In semantics, an ontology is an artefact which usually takes the form of a database on a computer and which contains a formalized record of what exists, often organized as a directed graph that allows the computer to infer facts, such as that a horse is an animal. It is in this sense that *ontology* is used here. Specifically, an ontology called WordNet (Fellbaum, 1998) is used here.[3]

WordNet is a large computational lexicon of English in which the meaning relations between words are encoded explicitly. The basic relation is synonymy: words which are deemed to be synonyms are grouped into objects called *synsets* (short for *synonyms sets*). The relations between words and synsets is many-to-many: a synset may contain multiple words and a word may be a member of multiple synsets. In effect, synsets represent the sense of a word as lexicographers know them – although in WordNet, the meaning of a word is often split up very finely

[3]The term *ontology* is used here in a broad sense. Strictly speaking, WordNet is not an ontology because it only records that subset of "what exists" which is lexicalized in English. It only passes for an ontology by proxy. That said, I believe it is still more appropriate to refer to WordNet as an ontology than as a thesaurus or a dictionary.

into a larger number of senses than is common in dictionaries.

The second most important relation in WordNet is hyponymy. This relation holds between synsets (as opposed to words) and records the fact that one synset (called the hyponym) is a special case of another synset (called the hypernym): a DEPUTY is a POLITICIAN, a POLITICIAN is a PERSON, a PERSON is an ANIMATE BEING, and so on. In the part of WordNet which contains nouns, all synsets are ultimately subsumed by a common parent called ENTITY. Hyponymy is the fundamental organizing principle of the noun lexicon in WordNet. It is very detailed and often strongly mirrors scientific taxonomies, as opposed to "folk taxonomies" – for example, HORSE is subsumed by ANIMAL not directly but by virtue of being an ODD-TOED UNGULATE, which is a HOOFED ANIMAL, which is a PLACENTAL MAMMAL, which is MAMMAL, which is a VERTEBRATE, which is a CHORDATE, which is eventually an ANIMAL.

Other prominent relations in WordNet include meronymy (something is part of something else), antonymy, entailment and a host of other relations which sometimes complement hyponymy by providing additional information about synsets, and sometimes replace hyponymy as the fundamental organizing principle in sections of the WordNet lexicon – for example, antonymy and not hyponymy is the main organizing relation among adjectives.

On the whole, WordNet is an attempt to explicate the internal structure of the mental lexicon – or of some aspects of the mental lexicon, at least. WordNet has been chosen for this work because of its wide lexical coverage and easy availability, and because previous work on selectional preferences has used it too. It is significant, however, that WordNet is more a product of reasoning and less a product of observing authentic language in use. This affects negatively its suitability to the task at hand, a point we will return to in chapter 7.

3.4 How to detect meaning clusters

Once a list of collocates has been extracted from a corpus, we want the computer to detect any clusters of related words on the list. How might we go about this? An obvious place to look for information about the relatedness of words is in hierarchically organized ontologies such as WordNet.

How can a lexical ontology be used to find clusters among a list of words? Essentially, there are two ways the task can be approached: one makes use of similarity measures, and the other – championed by Resnik – works by matching wordlists to categories in the ontology. In the former approach, a measure of semantic similarity (or semantic relatedness) between each pair of words is determined from WordNet using one of the many available techniques (Budanitsky and Hirst, 2006), and words whose similarity to each other exceeds a given threshold are declared to constitute a meaning cluster. For example, when processing a list of the direct objects of *follow*, the algorithm will compute a similarity measure between all pairs of words and determine that the words *direction*, *rule* and *instruction* are all similar to each other to a measure that exceeds a specified threshold. The three words would then be deemed to constitute a cluster.

This method has the disadvantage that it does not actually reveal why the words are related: it merely supplies a numerical score of relatedness. If we wish to name the reason for the similarity, we must avail of Resnik's approach in which we essentially try to "peg" a wordlist onto a category in the ontology – for example by declaring that the set *direction*, *rule*, *instruction* corresponds to the synonyms and hyponyms of the WordNet synset 'a message describing how something is to be done'.

3.5 Naming selectional preferences

It is important that we not only detect the existence of semantic clusters among a list of words but also name them, such as by reference to a WordNet synset.

The existence of a formal naming scheme will go a long way towards synonym comparison and towards the detection of translation equivalents. If we know that *sustain* has a selectional preference for INJURY, which is a synset in WordNet, we can look for other verbs that have a preference for the same synset and hopefully locate *suffer*. In a bilingual situation, if we are looking for a translation equivalent for the Irish verb *ól* 'drink' + TOBACCO, where TOBACCO is a specific synset in WordNet, we should be able to retrieve automatically that *drink* does not have a preference for this synset, while the less conventional translation *smoke* does.

3.6 Dealing with ambiguity

In this work, we wish to induce a word's selectional preferences by comparing a corpus-extracted list of its collocates to the categories existing in WordNet. Such induction process is complicated by polysemy. The noun *bullet* has two senses in WordNet (it is a member of two synsets): 'projectile fired from a gun' and 'high-speed train' but only one of them is evoked when *bullet* appears as the direct object of the verb *fire*. The induction mechanism has no way of knowing this, however, as it only has access to the textual traces of meaning. Both meanings of *bullet* must therefore be considered in the induction process. Although many words in the input set are usually polysemous in many different directions, most of the time the commonalities of meaning prevail and will sway the induction process in the right direction (Light and Greiff, 2002, p. 272).

3.7 Dealing with noise

Ambiguity has the potential to lead induction astray, as do phenomena such as type shifting and metaphorical usage. Another problem is "noise" which comes in the form of tagging errors (when *flying* is tagged as a verb instead of an adjective in *she saw flying planes*) and parsing errors (when the sequence *eating disorder* is misinterpreted as a verb followed by a direct object).

We can also count as noise the fact that we are ignoring some information during parsing. In this study, we are looking at words rather than constituents. In the sentence *they ate chocolate motorcycles* we extract not the whole noun phrase *chocolate motorcycles* but only the head noun *motorcycles*, thus depriving ourselves of the chance to deduce that the referent is edible. It may seem unwise to ignore everything but the lexical head, considering that selectional preferences do "hold of constituents and not simply lexical items" (Resnik, 1996, p. 48). However, it would increase the complexity of the task enormously if we were to consider constituents in their entirety and to infer their meaning compositionally. Most of the time, the semantic type of the constituent's referent is identical to the semantic type of the head's referent. Cases like *chocolate motorcycles* are exceptional.

Fortunately, the presence of any kind of noise is usually limited to a minority of examples. It is believed that a well-designed, balanced corpus contains an overwhelming proportion of "good" examples, that is, collocations which are not metonymic extensions, unconventional metaphors, or parsing errors.

3.8 Generalizing at the right level

The induction of selectional preferences from corpus data is basically a process of making generalizations. When given the set {*apple, berry, grape*}, we want to pronounce a generalization that unites them, such as the fact that they are all kinds of FRUIT. However, there is the question of which level in the taxonomy is the most appropriate for such generalizations. In WordNet, an APPLE is an EDIBLE FRUIT which is a FRUIT which is a FOOD which is a SOLID which is a MATTER which is a PHYSICAL ENTITY which is an ENTITY. Which of these levels is the appropriate one for generalizations? In the case of {*apple, berry, grape*} the answer appears straightforward: the lowest common hypernym is the appropriate level, which in this case is EDIBLE FRUIT. But what if *apple, berry, grape* is a subset of a larger set of words, such as the direct objects of *eat*: {*bread, apple, meat, cake, berry,*

grape, lunch}? Do we want the induction to reveal that some of the things that can be eaten are EDIBLE FRUIT? Or do we want to ignore this detail and simply state the larger generalization that the words denote kinds of FOOD? The answer depends on what the pivot word is. In the case of *eat*, we will probably be happy with the larger generalization (FOOD). In the case of the direct objects of *pick*, we probably want to isolate FRUIT as a distinct set, so the right answer here would be several distinct lexical sets including the set of FRUIT. After all, we eat fruit because it is food, not in addition to food – but we pick fruit in addition to other things we pick, not because it is food. As humans we know this and we use this knowledge effortlessly to decide whether FRUIT or FOOD is the appropriate level for generalizations in each case. The computer, however, does not have access to this knowledge. The lack of clarity as to the appropriate level of generalization is the most common source of misgeneralizations, as will be seen later. For example, when configured inappropriately, the system may over-generalize that the direct objects of *pick* are kinds of PHYSICAL ENTITY, failing to notice the salient subset FRUIT.

One possibility that seems to offer itself is to make generalizations at or near the basic naming level (Rosch, 1978). However, WordNet contains no explicit indication of which level in the hierarchy is the basic level. Even if it did, it is questionable whether that would be useful for our purposes because, as has already been shown, the appropriate level for generalizations is context-dependent.

3.9 Selectional preferences as predictions

An important application of the ability to name selectional preferences formally is the ability to make predictions. Once we have induced from several corpus-attested citations, such as *send a letter* and *send an e-mail*, that *send* has a selectional preference for the synset MESSAGE, then we can predict that the verb will be equally capable of taking as objects other words that denote the synset and its hyponyms,

such as *telegram* and *card*, even if those have actually not been attested in our corpus. The ability to make predictions is important for word-sense disambiguation and for natural-language generation. Collocational sets are typically open-ended: there is no end to the number of items that can qualify as HUMAN or FOOD. The only way to make predictions is to abstract away from extensionally defined collocational sets (with members listed explicitly) and to try and induce their intensional definitions.

An evaluation of the predictions made can help us recognize whether the generalization has been made at the right level. If the system has overgeneralized *pick* + {*apple, berry, grape*} as *pick* + FOOD, then it will predict incorrectly that we can also *pick*, for instance, *sausages* and *spaghetti*. Incorrect predictions like these suggest that the generalization needs to be made at a lower level. Crucially, the only way to evaluate whether the predictions are correct or not is human intuition, as the predicted collocations will not be attested in the corpus from which they have been induced.[4]

3.10 What kinds of selectional preferences are there?

It seems that a large number of selectional preferences can be explained by hypernymy. For example, the direct objects of *eat* overlap with the WordNet synset FOOD and its hyponyms. This way of looking at selectional preferences has been studied extensively by Resnik (1993) and his followers.

That is not the entire picture, however. Some selectional preferences seem to be better explained by meronymy. There are cases when the objects a verb prefers are meronyms of the same holonym but not hyponyms of the same hypernym. For example, a subset of the direct objects of *bury* are the nouns {*head, face, mouth, nose*} (as in *she buried her face in her hands*, *he buried his nose in her neck*). Importantly, the four nouns are not closely related in the hypernymy hierarchy: their

[4] Alternatively, the predictions could be verified in a larger corpus. However, if a larger corpus is available, then why not use it for the induction as well?

nearest common hypernym is the over-general OBJECT. A closer connection is provided by meronymy because the nouns denote HEAD and parts of it, and the fact that FACE, MOUTH and NOSE are parts of HEAD is encoded explicitly in WordNet. Therefore, if relevant generalizations are not to be missed, it is important to exploit other types of relations beside hyponymy. Additionally, it seems that there are two types of meronymy-based selectional preferences. One type is when the holonym is included in the preference and another type is when it is not. The direct objects of *bury* include meronyms of HEAD as well as *head* itself. In contrast, the indirect objects in the pattern *stab somebody in something* include various meronyms of BODY such as *chest* and *neck* but not *body* itself (we do not say *the killer stabbed the victim in her body*).

Large as WordNet is, it does not capture every potentially relevant fact. For example, WordNet does not tell us that HORSE and CAR are both used for transport, or that HELL and FEAR share a connotation of negativity. Such lexical relations belong in the realm of what are sometimes termed "non-classical" relations (Morris and Hirst, 2007). While WordNet cannot provide us with those, other databases can – such as ConceptNet[5] (Liu and Singh, 2004) for various relations including telic ones ("what things are for") and SentiWordNet[6] (Esuli and Sebastiani, 2006) for sentimentality scores (positive versus negative connotations).

Finally, there is the question whether we can expect that all selectional preferences will always correspond to some category or other in an ontology. If we come across a selectional preference that does not seem to fit any category in WordNet or in some other ontology, what is the significance of this finding? Does the ontology contain an omission and should it be corrected? This is a serious question with serious implications and we will return to it later. In the meantime, we will spend some time building a software tool which will allow us to study the relationship between selectional preferences and ontologies on practical examples. The tool, discussed in the next chapter, makes it possible for the researcher to cross-check a

[5] `http://web.media.mit.edu/~hugo/conceptnet/`

[6] `http://sentiwordnet.isti.cnr.it/`

collocational lexical set against various kinds of lexical sets derived from WordNet and based on hyponymy, meronymy and a number of other relations.

Chapter 4

The making of SenseMaker

All previous projects which have used a combination of corpora and WordNet to investigate selectional preferences, including Resnik's own work (Resnik, 1993), were limited to the hyponymy hierarchy of nouns, and to the preferences of verbs to take those nouns as their direct objects. Each occurrence of each noun in the object position counted as evidence that the verb has a preference for the synset (or synsets) the noun is a member of, as well as its hypernyms. Various statistical methods were then used to interpret the evidence and to induce which synsets, along with their hyponyms, constitute best matches for the linguistic evidence.

The present project makes use of a wider range of relations than hyponymy and is extendable with additional relations of any kind, even from ontologies other than WordNet. The challenge is that one does not want to have to update the induction algorithm each time a new relation is added. To remove this complexity, I have decided not to access WordNet directly and instead to convert the richly complex structure of WordNet into a fairly flat (but large) database of explicitly stated lexical sets – corresponding intuitively to "a set of words that denote kinds of animals", "a set of words that denote things that have blades", and so on. This chapter describes how this pre-processing step was done, and how the lexical sets thus obtained are then exploited.

Importantly, the design adopted here lends itself easily to extension. Because

the semantic data exists in the form of flat lexical sets (that is, sets of words rather than sets of synsets), new sets can always be added, even from sources that do not contain any explicit connections to WordNet synsets.

The induction performed by SenseMaker is less mathematically sophisticated than the algorithms developed by Resnik and his followers (Light and Greiff, 2002) but it has a wider scope in the sense that it can work with lexical sets other than just hypernymy.

4.1 Deriving hypernymy-based lexical sets

The hyponymy hierarchy is often referred to as the backbone of WordNet. This is true in particular of the noun synsets because all noun synsets are ultimately descendants of the synset ENTITY. This hierarchy is also what previous studies of selectional preferences have concentrated on.

The process for deriving lexical sets from the hyponymy hierarchy is intuitively very simple. We start with a synset, for example *fear, fearfulness, fright*, and wish to derive a lexical set containing words denoting kinds of FEAR. First, we acknowledge that the synonyms listed as members of the synset are also members of the lexical set. Then we include all the direct hyponyms of this synset, including the likes of *alarm, dismay, consternation* and *apprehension, apprehensiveness, dread*, and acknowledge that these words are also members of the lexical set. Then we expand our range one hyponymy level deeper and repeat the process. Once we reach the bottom of the hierarchy, we have collected all the words that denote kinds of FEAR, including the word *fear* itself and its synonyms.

I designed the lexical sets so as to contain an indication of the depth at which each word belongs to it – in effect, the lexical sets are sets of the two-tuples $<$*word, depth*$>$. The words *fear* and *fearfulness* are members of the set FEAR to a depth of 0, the words *alarm* and *apprehension* to a depth of 1, and so on. In all, the set FEAR contains 49 different words at depths ranging from 0 to 3. As will be seen

later, the notion of depth is significant in the induction stage.[1]

Last but not least, WordNet recognizes a relation of "instance" in addition to that of hyponymy. This relation holds between individual entities and general concepts, for example PARIS is an instance of CITY. This relation has been included in the definition of hyponymy for the purposes of lexical set derivation, so PARIS "is a" CITY in the same way as ALARM "is a" FEAR.

Some lexical sets obtained in this way are very large: in particular, the lexical set denoting kinds of ENTITY contains all the nouns of WordNet, while other lexical sets are very small: the set denoting kinds of PSALM contains only one word, *psalm* itself. In total, this method has produced 121,280 lexical sets – exactly as many as there are synsets in WordNet. Hypernymy-based lexical sets carry the title KINDS-OF in the database, for example KINDS-OF FEAR.

4.2 Deriving meronymy-based lexical sets

Arguably the second most important relation after hypernymy/hyponymy in WordNet is the meronymy/holonymy relation which holds between parts and the wholes they are contained in. Two types of lexical sets can be derived from this relation. The first is PARTS-OF, sets containing words that denote parts of something such as parts of BIRD: {*wing, beak, feather,* ...} and parts of BODY: {*head, foot, arm,* ...}. The other type is CONTAINERS-OF, sets of words that denote concepts that contain something, for example containers of TIP: {*knife, pen, flagpole,* ...} and containers of HANDLE: {*knife, door, saucepan,* ...}.

One question that needs to be answered before one sets about deriving lexical sets from the meronymy relation is whether meronymy is transitive. If A contains B and B contains C, does it follow that A contains C? Miller and Fellbaum (2007, p. 271) admit that this only follows "with qualification". While it is beyond doubt

[1] Synset depth here essentially acts as a substitute for synset specificity. The notion of specificity in WordNet remains undefined but depth seems to be a strong indicator of specificity (Vogel and Devitt, 2004).

that HOUSE contains DOOR and DOOR contains HANDLE, it appears slightly odd to state that HOUSE contains HANDLE. One intuitively hesitates to accept such statements because they "skip one step" and do not say "the whole truth". One might call such statements 'meronymously transitive'. Meronymously transitive statements are unlikely to occur in every-day discourse but that does not render them false. In fact, meronymously transitive statements are true in the sense that they truthfully describe the actual state of affairs. It only takes a moment's reflection to realize that houses do contain handles. The fact that this statement is not worth uttering does not detract from its truthfulness. Therefore, I have decided to treat meronymy as transitive during the derivation of lexical sets.

Another question concerns what I will call 'hyponymy expansion'. If A contains B and C is an A, then it would seem to follow that C also contains B. For example, if HUMAN contains FOOT and MIDGET is a HUMAN, then MIDGET also contains FOOT, even if this fact is not stated in WordNet explicitly. The example just stated is an example of hyponymy expansion on the holonym side: whatever the holonym contains, its hyponyms also contain. Hyponymy expansion can be done on the meronymy side, too: whatever the meronym is part of, its hyponyms are also part of. For example, if HAND contains FINGER and INDEX FINGER is a FINGER, then HAND also contains INDEX FINGER. The examples shown so far suggest that hyponymy expansion is a valid technique that yields truthful inferences, but there are other examples that suggest otherwise. For example, unchecked hyponymy expansion on the meronym side would force one to assert that HUMAN contains CROW'S FOOT (because HUMAN contains FOOT and CROW'S FOOT is a FOOT).

Why is this? One can think of meronymy as a relation that assigns attributes to concepts: the fact that A contains B is an attribute of A, and the fact that B is part of A is an attribute of B. Theoretically, one would expect that attributes should be inheritable down the hyponymy hierarchy without constraint, so that all the attributes of a hypernym are implicitly attributes of its hyponyms, too. But

things do not seem to work this way in real-world taxonomic hierarchies, WordNet included – at least not always. It seems that inherited attributes can be overridden by the hyponym but it is not clear when or why. Pending further investigation, I have decided to err on the side of caution and, rather than run the risk of producing false inferences, I have not performed hyponym expansion during the derivation of lexical sets from meronymy.

In total, the meronymy relation in WordNet has yielded 9,627 PARTS-OF sets and 20,405 CONTAINERS-OF sets.

4.3 Deriving lexical sets from other semantic relations

Besides hypernymy and meronymy, several additional semantic relations exist in WordNet which can also be exploited as sources of lexical sets. One of them is the relation of class membership which holds between noun synsets and specifies which concepts belong in which semantic class, for example the semantic class BIOLOGY contains the concepts CELL, MONAD, PHYLUM, SPECIES and others, the class MILITARY contains concepts such as GUN, COMBAT, DRILL, DEFENCE and BATTLE, and so on. This relation has yielded 630 lexical sets titled TERMS-FROM, for example TERMS-FROM BIOLOGY and TERMS-FROM MILITARY.

Another relation is entailment. This relation holds between verb synsets and specifies verbs that entail other verbs, such as OVERSLEEP, AWAKEN and DREAM which all entail SLEEP. In WordNet, entailment is deemed to hold between concepts A and B if it is acceptable to state that to do A, you must also be doing B (Miller and Fellbaum, 2007, p. 287) – for example, to *snore* you must also *sleep*. This relation has yielded 288 lexical sets titled TERMS-ENTAILING, including TERMS-ENTAILING SLEEP.

Finally, there are two relations in WordNet which pertain specifically to adjectives. One is the relation of similarity which unites adjective synsets with similar meanings. Adjective synsets with similar meanings are deemed to constitute clus-

ters, and inside each cluster one synset is deemed to be the cluster head while the other synsets are its satellites. For example, there is a similarity cluster around the head ABRIDGED which includes the satellites SHORTENED, HALF-LENGTH and POTTED. This relation has yielded 2,510 lexical sets titled SATELLITES-OF, such as SATELLITES-OF ABRIDGED.

The second relation which pertains specifically to adjectives is the relation of attributes. This relation links adjective synsets such as FAST and SLOW to noun synsets such as SPEED: in more general terms, it links values to attributes where the adjective synsets are the values and the noun synsets are the attributes. In many cases no more than two adjective synsets are linked in this way which are also antonyms of each other. This relation has yielded 320 lexical sets titled VALUES-OF, including VALUES-OF SPEED.

4.4 Inducing the selectional preferences

Once a database of lexical sets is in place and a list of corpus collocates is ready, the two need to be compared. Essentially, we are looking for overlaps between the collocate list and each of the lexical sets, trying to find sets that overlap with the collocate list to the highest degree. For instance, suppose that we have extracted from the corpus the following list of direct objects of some verb:

L = {*call, letter, message, vessel, ship*}

And we also have a database containing the following lexical sets:

COMMUNICATION = {*call, letter, message, missive, e-mail, communication*}

SHIP = {*vessel, ship, boat*}

POT = {*pot, vessel, pan*}

BODYPART = {*hand, foot, head, neck, finger*}

It is obvious that the sets COMMUNICATION and SHIP constitute good matches, the set POT constitutes a poor match, and the set BODYPART constitutes no match at all. The challenge now is to formalize this notion into an algorithm.

A simple way to start is to see which of the lexical sets in the database have the greatest overlap with the collocate set, measured in terms of the number of words. Typically, a large number of sets will be returned by this method, including many overgeneralizations. To cut down on the amount of overgeneralization, I have introduced the notion of *generality* into the induction process. Generality is related to the concept of depth in set membership, a concept which has been explained earlier in this chapter (4.1). Essentially, the membership of every word in a lexical set is qualified with a measure of its depth. The word *fear* is a member of the lexical set KINDS-OF FEAR at depth 0, of the set KINDS-OF EMOTION at depth 1, of the set KINDS-OF FEELING at depth 2, of the set KINDS-OF STATE at depth 3, and so on. During induction, generality is a numerical value which determines how far removed from depth 0 a lexical set may be if it is to be considered a match. If we specify generality to be, say, 0, then the overlap between a lexical set in the database and the collocate set must include at least one member at depth 0. For example, when faced with the collocate set {*fruit, nose, apple, berry, brain*}, the system will notice that there is a set in the database, namely FRUIT, which matches three members of the list, one of them at depth 0. Thus it concludes that FRUIT is the most valid generalization.

It appears that in most cases, the default setting of generality to 0 produces the most intuitively valid generalizations. In some cases when it does not, setting generality to 1 improves the intuitive validity of the results. A generality higher than 1 is almost never needed. Some real-world examples of the algorithm's performance will be shown in Chapter 6.

4.5 Displaying the results

SenseMaker works by comparing the word list supplied to it against all lexical sets in its database, and returning the most likely matches. The results take the form of a list of lexical sets (see figure 4.1) where each set is specified by its type (such as KINDS-OF, PARTS-OF) followed by an indicative word (such as KINDS-OF ANIMAL, PARTS-OF HOUSE) and the WordNet synset gloss. SenseMaker also shows the words from the list supplied which were matched against each lexical set, and a link to display other words from the same lexical sets (indicated with ellipsis: ...). In effect, these constitute predictions. For example, SenseMaker may conclude that from the list of the direct objects of *eat*: {*bread, apple, meat, cake, berry, grape, lunch*}, the three words {*apple, berry, grape*} match the lexical set KINDS-OF FRUIT. In that case, SenseMaker will provide a link to display all other words that belong to that lexical set, such as *cherry* and *banana*, thus making the prediction that these words may also appear as direct objects of *eat*, even though they have not been attested in the corpus.

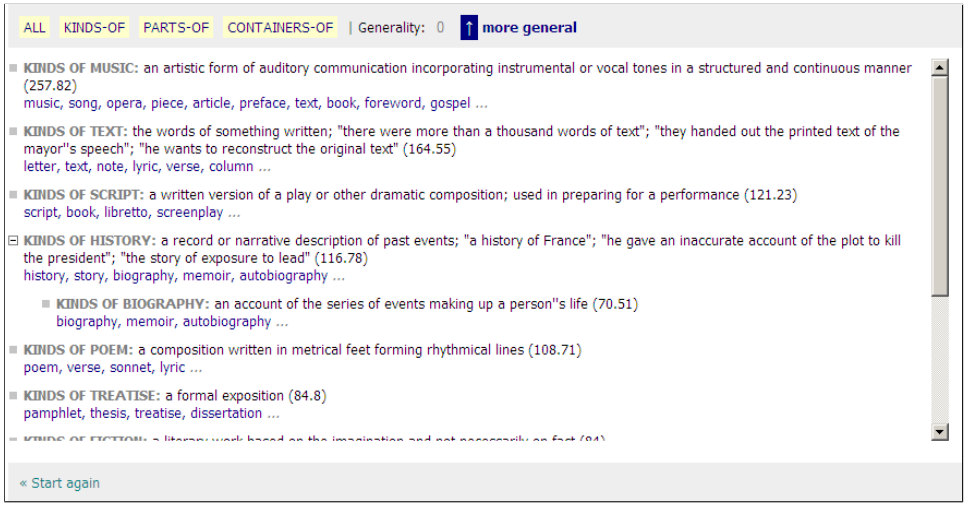

Figure 4.1: SenseMaker displays its generalizations for the 50 most salient direct objects of the verb *write*.

The lexical sets are listed in an order of the matched words' combined salience or frequency, if supplied, and if not, then they are ordered simply according to the number of words matched. As SenseMaker has access to WordNet's hypernymy and meronymy hierarchies, the listing is hierarchical. For example, if SenseMaker concludes that both KINDS-OF EDIBLE FRUIT and KINDS-OF FOOD are matches and if the synset EDIBLE FRUIT is a hyponym of the synset FOOD in WordNet, then SenseMaker will list the set KINDS-OF EDIBLE FRUIT underneath the set KINDS-OF FOOD, to indicate that it is a subset.

4.6 Using SenseMaker

The easiest way to use SenseMaker is to extract from a corpus a list of a certain number of the most frequent (or the most salient) collocates of some word, say the 100 most salient direct objects of *send*, then to paste the list into SenseMaker, and see if SenseMaker discovers any semantic clusters in it. Another possibility is to first subdivide the list manually into semantic clusters which the human user finds intuitively valid, and then paste those clusters into SenseMaker one at a time and see if it finds a match for them in its database of lexical sets. In each case, the final (and optional) step is to review the predictions SenseMaker makes.

Chapter 5

Case studies

This chapter presents five case studies in which I used SenseMaker to try to match the selectional preferences of 5 randomly selected words with the lexical sets I had derived from WordNet in the previous chapter.

The 5 words were chosen at random from a list of over 6,000 most frequent lemmas extracted from the British National Corpus[1]. The only conditions were that they are common words (words which occur fewer than 2,000 times in the corpus were ignored) denoting predicates (i.e., verbs and adjectives) and that they are distributed evenly in terms of frequency: 2 words were chosen form the frequency band 2,000–6,000 and three from the frequency band >6,000. I decided to choose the words randomly in order to avoid any temptation to select words that appear likely to yield "interesting" selectional preferences. My study is motivated by the desire to study the normal, conventional use of language. What linguists consider interesting is not necessarily normal or conventional.

For each word, I chose one syntactical relation which I believed to be most central to the predicate. Typically, the relation is the direct object (for verbs) or the noun modified (for adjectives). Table 5.1 gives a listing of the 5 words and syntactical relations chosen. Then, using the Sketch Engine, I extracted a list of the 100 most salient arguments that stand in this syntactic relation to the predicate.

[1] http://kilgarriff.co.uk/bnc-readme.html

This list is an extensionally defined lexical set whose members are the lexical traces of the predicate's selectional preference for the given syntactical relational. Finally, I used SenseMaker to try and find matches for the set (or more often, for subsets of it) in the database of lexical sets I had derived from WordNet.

The case studies are intended as exploratory, they are not designed to yield a quantitative measure of SenseMaker's performance. A larger-scale study would be necessary for that.

The outcomes of the tests are summarized in the rest of this chapter. These provide empirical data for the conclusions drawn in the next chapter. Example sentences in this chapter are authentic and have been extracted from the British National Corpus.

Table 5.1: The words and syntactical relations chosen for the test.

BNC frequency	Word	Part of speech	Syntactical relation
32,675	*live*	verb	subject
9,266	*push*	verb	direct object
6,107	*immediate*	adjective	noun modified
3,973	*willing*	adjective	subject
2,205	*cancel*	verb	direct object

5.1 Case study 1: subjects of *live*

The top 100 subjects of the verb *live* are straightforward as the large majority of them refer to living things, the most prominent subset being people. SenseMaker matches them correctly to the synset ORGANISM and its hyponym PERSON.

There are cases when the subject of *live* refers not to a person but to a group of people such as *family*, *population* and *generation*. The nearest match for these in WordNet is the synset SOCIAL GROUP but this does not include *population* and *generation*. The nearest hypernym of SOCIAL GROUP that does include *population*

and *generation* is GROUP, but this is an overgeneralization as it subsumes groups of non-living things as well. Another candidate is the synset PEOPLE, glossed as '(plural) any group of human beings (men or women or children) collectively' – this does include *population* and *generation* but does not include many others such as *family* and *household*. Neither of the meronymy-derived lexical sets is of any help, either: the synset PERSON is a meronym of PEOPLE and its hyponyms are meronyms of a handful of other synsets (including PARENT, CHILD and SIBLING being meronyms of FAMILY), but there is no way in WordNet to infer that, for example, HOUSEHOLD contains PERSON. It turns out that WordNet does not have a category that would subsume all kinds of groups of people.

Surprisingly, no metaphorical uses of *live* (such as *the legacy lives on*) are frequent enough in the British National Corpus to have made it to the list of the 100 most salient subjects of the verb.

5.2 Case study 2: direct objects of *push*

It comes as no surprise that the things we most often talk about *pushing* are physical objects, and SenseMaker matches them to the WordNet synset PHYSICAL ENTITY. A prominent subset is BODY PART (e.g. *Barbara pushed her hands into her overall pocket*), which SenseMaker matches to the meronymy-derived set PARTS-OF BODY. Apart from that, SenseMaker also reveals that we can push kinds of BOUNDARY (*boundary, limit, frontier*) and kinds of IDEA or THOUGHT (*idea, thought, claim, argument*).

The case of *pushing a boundary* deserves further investigation. SenseMaker matches this to the synset BOUNDARY glossed as 'the line or plane indicating the limit or extent of something', but this makes some incorrect predictions such as *push a face*. The construction *to push someone's face somewhere* is of course felicitous, but not as an instance of *push a boundary*. In WordNet, *face* is subsumed under BOUNDARY in its sense of 'a surface forming part of the outside of an object',

but that is not the sense evoked in *pushing someone's face somewhere*. The sense evoked in *pushing someone's face somewhere* is that of *face* as BODY PART. So the generalization <*push*, dir-obj, BOUNDARY> is in fact not completely correct, even though it appears intuitively valid at first scrutiny. Some boundaries can be pushed and some cannot, but WordNet does not provide a way of distinguishing which is which.

5.3 Case study 3: nouns modified by *immediate*

The adjective *immediate* conveys roughly the idea that two things are adjacent, without anything intervening between them. This broadly defined meaning is exploited copiously in discourse to connect physical objects with their surroundings (*in the immediate vicinity of the airport*), to connect events to events that follow (*their immediate reaction was negative*) or to events that precede (*the immediate cause of death was in the stomach*), to connect people to other people (*he reversed the policies of his immediate predecessor*), and a host of other meanings which resemble one of the above to varying degrees.

The question is, can any generalizations at all be made about the selectional preferences of *immediate*, seeing that its range is so wide? We could take the easy way and conclude that *immediate* may modify ANYTHING (or whatever it is we call the uppermost category in our ontology) but unfortunately, that would result in many false predictions. Clearly, *immediate* cannot normally combine with *sofa*, *foot* or *driver*. It is obvious that *immediate* is, after all, quite particular about what it qualifies – in other words, it does have a selectional preference.

If we wish to find and name *immediate*'s selectional preferences, we can start with the three categories mentioned above (SURROUNDINGS, EVENTS and PEOPLE) and see if SenseMaker finds any matches for them its database of lexical sets derived from WordNet. Let us start with SURROUNDINGS. This collocational sets includes nouns like *vicinity, surroundings, neighbourhood, environs* and *hinter-*

land. SenseMaker finds that the most likely candidates are the synsets GEOGRAPHICAL AREA and SECTION (the latter is glossed as 'a distinct region or subdivision of a territorial or political area or community or group of people') but neither contains all the words (GEOGRAPHICAL AREA does not include *locality*, SECTION does not include *hinterland*) and both make some incorrect predictions: we do not normally say *the immediate town* or *the immediate meadow* even though they are both kinds of GEOGRAPHICAL AREA.

Very similar results obtain when we use SenseMaker to find matches for the other categories mentioned, that is, nouns that denote EVENTS and nouns that denote PEOPLE. We end up with synsets which appear intuitively correct but turn out to be too general because they make incorrect predictions.

Interestingly, SenseMaker matches some of the words from the collocational sets with the lexical set TERMS-FROM LAW. The words matched include *interest, relief, cause, effect, action, answer, use* and *release*. While not all of them are specifically law terms, not even when used in combination with *immediate*, this does correctly reveal that *immediate* belongs in a formal register.

5.4 Case study 4: subjects of *willing*

The subjects of the adjective *willing* are those nouns that the adjective stands in a predicative relation to, as in *more people might be willing to adopt animals* and *the few people willing to brave the property market*, but not when the noun is modified by the adjective in the same noun phrase, such as *willing helpers*.

Virtually all subjects of *willing* refer to humans and SenseMaker correctly matches them with the synset PERSON. As was the case with the verb *live* and its subjects, we see an extension here from PERSON to GROUP OF PEOPLE: not only *friend*, *teacher* and *minister* are *willing* to do things but also *government*, *country*, *bank* and *family*. And, as with the subjects of *live*, SenseMaker is having difficulty finding an intuitively valid match for this concept in WordNet because

there is no sufficiently general category in WordNet to capture the idea of "a group of people".

An interesting point is that the kinds of groups that appear as subjects of *willing* are slightly different from those that appear as subjects of *live*. We can say that *a government is willing to do something* but it is odd to say that *a government lives* (e.g. *in a nice house*). Even though the generalization GROUP OF PEOPLE appears intuitively valid, it is not accurate enough as it does not distinguish between the kinds of groups which can *live* and those which can be *willing*.

5.5 Case study 5: direct objects of *cancel*

SenseMaker reveals that most of the things we can *cancel* are kinds of EVENT. This is intuitively valid, but how precise is it? Some of the predictions it makes include events like *occurrence*, *incident* and *crash*. It is odd to say that somebody *cancels the occurrence of something* or that someone *cancels an incident*. Upon reflection, it turns out that one can only *cancel* events that have been pre-planned, such as *meeting*, *wedding* and *concert*. Therefore, for an event to qualify as a direct object of *cancel*, it has to be capable of being pre-planned. No such category exists in WordNet but some prominent subsets do, such as MEETING, JOURNEY and SHOW.

There is often an implication that if one cancels an event, the event has not happened yet. For example, cancelling a wedding normally implies that the wedding hasn't happened yet. But you can also cancel things with are already in progress such as *subscription*, *contract* and *registration*. These are not EVENTS but STATES or more precisely STANDING ARRANGEMENTS. Subscriptions and contracts are arrangements which usually last a long time and can be made to cease to exist (that is, cancelled) while they are in progress. Again, there is no such category in WordNet but some of its prominent subsets are, including CONTRACT and ARRANGEMENT, which SenseMaker reveals.

There is some conceptual overlap between the two types of things one can

cancel. For example, when you book a trip with a travel agent and then you *cancel* the booking, you are at the same time cancelling the standing arrangement that exists between you and the travel agent, as well as the pre-planned event of going on the trip. Additionally, there are cases of metonymic extension which confuse SenseMaker, the most striking of which is *milk*. In the phrase *cancel the milk*, the *milk* does not of course refer to the white liquid that comes out of a cow's udder, it refers to the standing arrangement of having one's milk delivered to the house every morning. Thus *cancel the milk* belongs in the same group as *cancel the subscription* and *cancel the contract* but SenseMaker does not group it as such because WordNet does not contain the sense of *milk* as 'arrangement to have milk delivered'.

5.6 Summary of the case studies

The results of the case studies are best summarized as inconclusive: the generalizations SenseMaker finds are sometimes correct and sometimes not. In particular, the generalizations found by SenseMaker are often too granular: where a human senses a single category, such as SURROUNDINGS, SenseMaker finds several, such as GEOGRAPHICAL AREA and SECTION.[2]

In other words, SenseMaker misses some relevant generalizations. Why is this? There are two possible reasons: the induction algorithm is inaccurate, or the ontology used is inaccurate. While not discounting the former, the next chapter (Chapter 6) will concentrate on the latter with a discussion of the relationship between selectional preferences and ontologies.

Interestingly, previous attempts to design a mechanism for inducing selectional preferences from corpora have been similarly inconclusive, as summarized in Light and Greiff (2002), even when they used (earlier versions of) the same ontology (WordNet). It appears that automatic selectional preference induction is bound

[2]This may not necessarily be a problem, given that underspecification is a useful tool for some applications.

to be subject to the same limitations as, for example, automatic part-of-speech tagging: not completely accurate, but accurate enough for some purposes provided one understands its limitations.

One purpose for which automatic selectional preference induction can be useful is to guide a human researcher's attention. A lexicographer who wishes to describe the selectional preferences of a word may find it useful to see how a tool like SenseMaker clusters the corpus evidence, as a first step. SenseMaker can also occasionally draw attention to less obvious generalizations which a human might miss, such as the fact that a significant category of the direct objects of *push* is BODY PART. A more detailed discussion of the potential applications follows in Chapter 7.

Chapter 6

Selectional preferences and ontologies

The single most important finding to conclude from the case studies is that selectional preferences are hard to "pin down". We have seen that what seems like a single concept, for instance GROUP OF PEOPLE, may in fact have several potential matches in WordNet, none of which is completely satisfactory.

This is not just a problem with WordNet. Even if we undertake to do the work manually and try to group the collocates of some word, say *immediate*, into semantically coherent clusters, we soon run into cases of indeterminacy. True, there are always a number of obvious cases: *vicinity* is a GEOGRAPHICAL AREA, *surrender* is an EVENT, *danger* is a STATE, and so on. But what is *impression* in a sentence such as *the immediate impression was good*? Is it an EVENT, is it a STATE, is it both, or is it a different category altogether?

This brings us to two important observations about selectional preferences. The first observation is that the semantic types which constitute a word's selectional preference are not mutually exclusive. Consider the verb *drink* and its direct objects. Some of its objects are LIQUID, some are CONTAINER (OF LIQUID) and some are QUANTITY (OF LIQUID). But which one of those is evoked in the sentence *he drank a bottle of beer*? Is the intended meaning that of QUANTITY ('this

is how much beer he drank'), or is that of CONTAINER ('this is what he drank beer out of')? Most of the time, the only reasonable answer is that it doesn't matter: the sentence *he drank a bottle of beer* is licensed not by one but two of its selectional preferences: <*drink*, obj, CONTAINER> and <*drink*, obj, QUANTITY> at the same time.

The second observation is that the semantic types which constitute a word's selectional preference seem to behave like fuzzy sets which are organized around a prototype and where a lexical item's membership is a matter of degree. For example, *impression* is a less prototypical EVENT than *surrender* (but still an EVENT enough to afford modification by *immediate*).

All this fuzziness demonstrated by selectional preferences stands in sharp contrast to the way ontologies like WordNet are organized. Fuzziness and prototypes do not exist in WordNet. Every word either is or is not a member of a category, and most WordNet users understand the categories to be mutually exclusive in the sense that a word in a sentence always realizes only one of its senses, never several at the same time.

Does this mean that selectional preferences are fundamentally irreconcilable with ontologies, and the effort should be abandoned? In spite of the evidence, the answer cannot be a resounding yes. For one thing, it can be shown that the marriage of selectional preferences and ontologies is useful for some purposes, even if the correspondence is not completely accurate. The practical uses of this will be dealt with in the next chapter. But more importantly, there is something intellectually satisfying about the generalization that we *cancel* EVENTS, in spite of the fact that not all EVENTS can be *cancelled* (*thunderstorms* and *car accidents*, for example). There appears to be an intuitive validity to such overgeneralizations, even if they are inaccurate. So before we reject ontologies altogether, some time should be spent reflecting on whether the semantic types embodied in selectional preferences are in fact fuzzy or not.

6.1 Are selectional preferences fuzzy?

The fact that the phenomenon of selectional preference can be observed in authentic language usage at all is, presumably, an outward expression of the categories that exist in the human mind. If humans are finding it acceptable to utter *their immediate surrender* as well as *their immediate impression*, then there must be some cognitive category in their mind that unites *surrender* and *impression* as members of the same type.

If we want to understand selectional preferences, we cannot avoid dealing with the phenomenon of mental categorization. Mental categorization is the process of assigning individuals to categories. When you come across an entity in the real world, be it something concrete such as a physical object you are looking at or something abstract such as an emotion you are feeling, you instinctively assign it to some category or other: you conclude that the object you are looking at is a bird, or that the emotion you are feeling is fear. Humans do this because it helps them think about the entity and to communicate the entity to other humans by means of language.

One way mental categories are demonstrated outwardly is by lexical choices. For example, if we believe that a *sparrow* and a *robin* are subtypes of the same type, we can express that by calling them by the same word, in this case *bird*. Another way in which mental categories reveal themselves to external observation is through selectional preferences: if we believe that *surrender* and *impression* are (in some way) members of the same type, we can express that by allowing them to be modified by the same adjective, in this case *immediate*.

Mental categorization is an active area of research and one of its central questions is whether mental categories are fuzzy or not. There is a long-standing tradition in Western thought to treat mental categories as crisp (that is, not fuzzy), with well-defined boundaries and not subject to prototypicality. This way of thinking goes as far back as Aristotle but is nonetheless subject to attack because the boundaries of concepts like HAPPINESS, FRUIT or COUNTRY are rather hard to formulate

precisely. A demonstration of the difficulty to find crisp boundaries for lexical concepts is Ludwig Wittgenstein's attempt to define the concept GAME (Wittgenstein, 1953). He concludes that there is no property common to all instances of GAME and instead redefines mental categories as "family resemblances".

A counter-attack on this position is Wierzbicka (1990) who claims to have found such common properties of GAME. Her point is that crisp boundaries for lexical concepts can actually be found, only it requires tremendous intellectual effort.

The same question pertains to mental categories as they are revealed by selectional preferences. Are they fuzzy or crisp? In other words, is it possible, at least in principle, to list all the necessary and sufficient conditions that completely define a collocational set, such as the set of things that can be modified by *immediate*? We can start answering the question by taking one or two collocational sets and trying to actually find and list those necessary and sufficient conditions. Let's return to the example of *drive* versus *ride* and their direct objects. We *drive* cars and trucks and buses but we *ride* motorcycles, bicycles, and horses. Can we find criteria that clearly demarcates the two sets form each other? What exactly distinguishes drivable objects from ridable ones? Is it the number of wheels? Is it the way the wheels are arranged? Is it the way you occupy the vehicle? Is it the presence of a steering wheel versus a handlebar?

It seems intriguingly difficult to decide which criterion, or which combination of criteria, demarcates the boundary. It appears as though each of those criteria merely contributes a certain amount towards deciding whether a concept is drivable or rideable. Difficult as it is to find crisp boundaries for the semantic types involved in selectional preferences, this does not imply that it is impossible. One thing is clear, however: even if the semantic types in selectional preferences are crisp, current ontologies, including WordNet, do not reflect them accurately.

6.2 Towards an ontology motivated by selectional preferences

The reason why WordNet does not reflect selectional semantic types accurately is because it was never designed to. Essentially, WordNet was compiled by asking informants questions such as *do you agree that a car accident is an event? do you agree that to be snoring, you must also be sleeping?* and so on (Miller and Fellbaum, 2007, p. 270ff) (in the case of WordNet the informants are mostly the compilers themselves but that is beside the point). These questions make an appeal to the informant's introspection and the answers to them are a product of reasoning. Thus, it is hoped, the internal organization of the mental lexicon is revealed.

But there is another way to reveal the internal organization of the mental lexicon, and that is to observe selectional preferences in action. We observe that humans combine *immediate* with *surrender* in the same way as they combine *immediate* with *impression*, and thus we can conclude that *surrender* and *impression* belong in the same category. Unlike the question-and-answer approach of traditional ontologies, this approach bypasses the informant's explicit reasoning capacity and instead focuses on his or her instinctive language use. Thus, it is a more direct window onto the mental lexicon. It is, after all, a well-known fact in empirical linguistics that people do not always use language in the way they claim they use it. Therefore, it is not unexpected that there should be a discrepancy between what exists in the mental lexicon and what people claim exists in it when prompted.

Having recognized this, the next obvious question is to ask, what should an ontology actually look like if it were to reflect accurately the semantic types involved in selectional preferences? It would certainly be much more language-specific than WordNet. But besides that, what would its organizing principles be? Should it be a bundle of Wittgensteinian family resemblances? Or should it be organized more or less hierarchically like current ontologies? And if so, how would it account for the phenomena of fuzziness and prototypicality?

To my knowledge, the only attempt to create an ontology that takes selectional preferences into account is the ongoing Corpus Pattern Analysis (CPA) project[1] conducted by Patrick Hanks at Masaryk University (Hanks and Pustejovsky, 2005). This project starts with an empty ontology called the Brandeis Semantic Ontology (Pustejovsky *et al.*, 2006). The ontology is organized according to traditional principles in the sense that it is a subsumption hierarchy and consists of semantic types such as HUMAN and EVENT. As part of the CPA project, the ontology is being populated manually with lexical items whose membership is qualified numerically and by context. For example, the word *meeting* is a member of the category EVENT to a degree of 0.12 when it is the object of *attend*, to a degree of 0.04 when it is the object of *hold*, and so on. The numerical values are derived from occurrence counts in a corpus. The ontology is organized the way it is in order to account for a phenomenon called "shimmering" (Hanks and Ježek, 2008): as a category moves from one context to another, some lexical items drop out and others join in. For example, as the category EVENT moves from *attend* to *hold*, some members drop out (you can *attend school*, but you cannot *hold* one) and others join (you cannot *attend a referendum* but you can *hold* one).

The CPA ontology is being populated manually and it will be some time before it becomes comparable to WordNet in terms of lexical coverage. Until then, practical applications which try to work with selectional preferences will have to work with ontologies which have not been designed for this purpose. Luckily, there are areas of application where the imprecision of current ontologies is not an issue, as the following chapter will illustrate.

[1] http://nlp.fi.muni.cz/projekty/cpa/

Chapter 7

Applications and further research

The previous chapter mentioned that pegging selectional preferences onto an ontology is useful, even if sometimes inaccurate. This chapter will expand on that point, by way of concluding the dissertation on a positive note. If we know that a verb has a selectional preference for EVENT, then that gives us a piece of information we can use, even if some kinds of EVENT are actually not included in the selectional preference. This chapter will suggest how the method studied in this dissertation could be exploited further for practical applications and for linguistic research.

7.1 Practical applications

A worthwhile project would be to compile a database of the selectional preferences of, say, English verbs. The database would be compiled semi-manually with the help of a tool like SenseMaker and would name the verbs' selectional preferences by reference to an ontology like WordNet. Such a database would be a valuable resource for practical applications in natural language processing (NLP) as well as for further research into selectional preferences themselves.

7.1.1 Applications for NLP

The benefits of having machine-readable knowledge of selectional preferences for natural language processing are associated with parsing, word-sense disambiguation and voice recognition, to mention just a few (see section 2.9). In many cases, the imprecision mentioned in the previous chapter does not cause difficulties. For example, if we wish to disambiguate the sense of *push* in the sentence the *we pushed the frontier of what's possible again* between its literal and its metaphorical sense, we only need to know that *frontier* is a BOUNDARY to conclude that the sense evoked here is the metaphorical one. It does not matter much that some BOUNDARIES, such as *surface* or *hairline*, cannot occur as indirect objects of *push* in this sense, as they are unlikely to occur in the literal sense either.

7.1.2 Cross-linguistic applications

An area touched on several times in this dissertation is the relevance of selectional preferences to the study of cross-linguistic equivalence. As has been demonstrated, a pair of words in two languages may seem like direct equivalents of each other at first sight, but will often differ slightly in their selectional preferences. An example is the English–Czech pair *subscribe* and *předplatit*. They differ in that you can *subscribe* to IDEOLOGIES in English but you cannot *předplatit* them in Czech.

It is achievable to use a multilingual ontology such as EuroWordNet[1] (Vossen, 1998) to build a database of selectional preferences in several languages which facilitates cross-linguistic comparison. EuroWordNet contains WordNet-like ontologies in several European languages, while equivalence between synsets in the various languages is indicated by links to an interlingual index. With these links, it is possible to induce automatically or semi-automatically that while *subscribe* has a preference for the synset IDEOLOGY in the English part of EuroWordNet, *předplatit* does not have a preference for its equivalent in the Czech part.

[1] `http://www.illc.uva.nl/EuroWordNet/`

7.2 Applications for research

The existence of a database of selectional preferences would open the door to further research into the phenomenon of selectional preference itself. The rest of this chapter will review some of the options.

7.2.1 The extensibility of selectional preferences

One question worth investigating is whether the extensibility of selectional preferences is regular in any way. In some cases, a predicate with a preference for HUMAN can also take ORGANIZATION, but there are predicates which cannot be extended in this way, as in (1) and (2).

(1) a. The president said so yesterday.
    b. The government said so yesterday.

(2) a. The president ate his dinner.
    b. #The government ate its dinner.

There are many such patterns of selectional preference extension, including ORGANISM extending into a BODY PART of that organism, EVENT extending into the PLACE where the event happens, LIQUID extending into a CONTAINER of the liquid. A statistical analysis of a database of selectional preferences could reveal whether there are any regularities in this behaviour, and whether they can be used for predictions. As an example, we could query the database for predicates that prefer both HUMAN and ORGANIZATION and contrast them with predicates which only prefer HUMAN.

7.2.2 From selectional preferences to patterns

A second area worth investigating is how selectional preferences co-occur. For example, we already know that *send* has a preference for MESSAGE (*he sent the letter yesterday*) and EMOTION (*it sent shivers down her spine*). That is not the whole

story, however. When *send* has an EMOTION for its object, it is invariably accompanied by an adverbial phrase headed by *down* where the complement is BODY PART. The whole pattern is '*send* EMOTION *down* BODY PART'. It is a pattern which combines elements of semantics (the semantic types EMOTION and BODY PART), syntax (EMOTION is the direct object, *down* BODY PART is an adverbial) and lexis (the words *send* and *down* are compulsory).

A statistical analysis of the co-occurrence of selectional preferences could reveal the existence of such patterns. If we had a database of selectional preferences which retains links to the corpus sentences from which they were induced, we could query it for pairs (or triples, or n-tuples) or selectional preferences which co-occur on the same token frequently enough or significantly enough. A statistically significant co-occurrence of (3-a) and (3-b) would provide evidence for the existence of the pattern (3-c), and so on.

(3) a. <*send*, obj, EMOTION>

b. <*send*, pp-*down*, BODY PART>

c. *send* EMOTION *down* BODY PART
(e.g. *it sent shivers down her spine*)

(4) a. <*bury*, obj, BODY PART>

b. <*bury*, pp-*in*, BODY PART>

c. *bury* BODY PART *in* BODY PART
(e.g. *he buried his nose in her neck*)

7.2.3 Abstracting away from individual words

A third area for investigation is to look at words which share a selectional preference and see what they have in common. This might help to discover new synonyms (*fix* has a synonym in *repair* when DEVICE is the object). But more importantly, words which share a preference for the same semantic type may themselves be of the same or related semantic type, as in (5) where all the verbs are of the type

EAT.

(5) a. <*eat*, obj, FOOD>

b. <*devour*, obj, FOOD>

c. <*nibble*, pp-*at*, FOOD>

It is possible that a further investigation of this might reveal interesting insights into the organization of the mental lexicon.

Chapter 8

Conclusion

All previous work which attempted to induce selectional preferences automatically from a corpus using an ontology such as WordNet, involved the unspoken assumption that there is an unproblematic correspondence between the mental categories embodied in the ontology on the one hand, and the mental categories revealed by selectional preferences on the other hand. This dissertation has revealed that this is not the case, and has discussed why.

However, the lack of correspondence between selectional preferences and ontologies is not as dramatic as to preclude the automatic induction of selectional preferences totally. This dissertation has demonstrated that selectional preferences induced from corpora, even if not completely accurate, can be usefully exploited for a number of applications.

Bibliography

Beattie, S.; Fellermayer, M.; Ohneis-Borzacchiello, E. (eds.) (2004). *Collins German Dictionary*. Glasgow: HarperCollins, 5th edition

Budanitsky, A.; Hirst, G. (2006). 'Evaluating WordNet-based measures of lexical semantic relatedness'. *Computational Linguistics*, 32

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press

Esuli, A.; Sebastiani, F. (2006). 'SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining'. In *Proceedings of Language Resources and Evaluation Conference 2006*. Genoa: European Language Resources Association

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press

Hanks, P. (2008). 'Mapping meaning onto use: a Pattern Dictionary of English Verbs'. Talk given to the American Association for Corpus Linguistics

Hanks, P.; Ježek, E. (2008). 'Shimmering lexical sets'. In *Proceedings of the 13th Euralex International Congress*. Barcelona: Universitat Pompeu Fabra

Hanks, P.; Pustejovsky, J. (2005). 'A Pattern Dictionary for Natural Language Processing'. *Revue française de linguistique appliquée*, 10(2):63–82

Katz, J. J.; Postal, P. M. (1964). *An Integrated Theory of Linguistic Descriptions*. Cambridge, MA: MIT Press

Kilgarriff, A.; Rychlý, P.; Smrž, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings of the 11th Euralex International Congress*. Lorient: Université de Bretagne Sud

Light, M.; Greiff, W. (2002). 'Statistical models for the induction and use of selectional preferences'. *Cognitive Science*, 26:269–281

Liu, H.; Singh, P. (2004). 'ConceptNet — a practical commonsense reasoning tool-kit'. *BT Technology Journal*, 22(4):211–226

McCawley, J. D. (1976). *Grammar and Meaning: Papers of Syntactic and Semantic Topics*. London: Academic Press, corrected edition

Miller, G. A.; Fellbaum, C. (2007). 'Semantic networks of English'. In Hanks, P. (ed.), *Lexicology: Critical Concepts in Linguistics*, volume 6. Oxford: Routledge

Morris, J.; Hirst, G. (2007). 'Non-classical lexical semantic relations'. In Hanks, P. (ed.), *Lexicology: Critical Concepts in Linguistics*, volume 6. Oxford: Routledge

Pustejovsky, J. (1993). 'Type coercion and lexical selection'. In Pustejovsky, J. (ed.), *Semantics and the Lexicon*, p. 73–94. Dordrecht: Kluwer

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press

Pustejovsky, J.; Havasi, C.; Sauri, R.; Hanks, P.; Littman, J.; Rumshisky, A.; Castano, J.; Verhagen, M. (2006). 'Towards a Generative Lexical Resource: The Brandeis Semantic Ontology'. In *Proceedings of Language Resources and Evaluation Conference 2006*. Genoa: European Language Resources Association

Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania

Resnik, P. S. (1996). 'Selectional constraints: An information-theoretic model and its computational realization'. *Cognition*, 61:127–159

Rosch, E. (1978). 'Principles of Categorization'. In Rosch, E.; Lloyd, B. B. (eds.), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates

Rundell, M. (ed.) (2007). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan, 2nd edition

Seuren, P. A. M. (1985). *Discourse Semantics*. Oxford: Blackwell

Soehn, J.-P. (2005). 'Selectional restrictions in HPSG: I'll eat my hat!' In Müller, S. (ed.), *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*. University of Lisbon

Vogel, C.; Devitt, A. (2004). 'The Topology of WordNet: Some Metrics'. In *Second International Wordnet Conference 2003*. Brno, Czech Republic: Masaryk University

Vossen, P. (ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer

Wierzbicka, A. (1990). '"Prototypes save": on the uses and abuses of the notion of "prototype" in linguistics and related fields'. In Tsohatzidis, S. L. (ed.), *Meanings and Prototypes: Studies in Linguistic Categorization*. London: Routledge

Wittgenstein, L. (1953). *Philosophical Investigations*, chapter Family Resemblances. Oxford: Blackwell