# Translation-based Word Sense Disambiguation

*Gunn Inger Lyse*

*Dissertation presented in partial fulfilment*
*of the requirements for the degree philosophiae doctor (PhD)*


Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
2011

# ABSTRACT

This thesis investigates the use of the translation-based Mirrors method (Dyvik, 2005, *inter alia*) for Word Sense Disambiguation (WSD) for Norwegian. Word Sense Disambiguation is the process of determining the relevant sense of an ambiguous word in context automatically. Automated WSD is relevant for Natural Language Processing systems such as machine translation (MT), information retrieval, information extraction and content analysis.

The most successful WSD approaches to date are so-called supervised *machine learning* (ML) techniques, in which the system 'learns' the contextual characteristics of each sense from a training corpus that contains concrete examples of contexts in which a word sense typically occurs. This approach suffers from a knowledge acquisition problem since word senses are not overtly available in corpus text. First, we therefore need a sense inventory which is computationally tractable. Subjectively defined sense distinctions have been the norm in WSD research (especially the Princeton WordNet, Fellbaum, 1998). But WSD studies increasingly show that the WordNet senses are too fine-grained for efficient WSD, which has made WordNet less attractive for machine-learned WSD. Ide and Wilks (2006) recommend instead to approximate word senses by way of cross-lingual sense definitions. Second, we need a method for sense-tagging context examples with the relevant sense given the context. Preparing such sense-tagged training corpora manually is costly and time-consuming, in particular because statistical methods require large amounts of training examples, and automated methods are therefore desirable.

This thesis introduces an experimental lexical knowledge source which derives word senses and relations between word senses on the basis of translational correspondences in a parallel corpus, resulting in a structured semantic network (Dyvik, 2009). The Mirrors method is applicable for any language pair for which a parallel corpus and word alignment is available. The appeal of the Mirrors method and its translational basis for lexical semantics is that it offers an objective and consistent—and hence, testable—criterion, as opposed to the traditional

subjective judgements in lexicon classification (cf. the Princeton WordNet). But due to the lack of intersubjective "gold standards" for lexical semantics, it is not an easy task to evaluate the Mirrors method.

The main research question of this thesis may thus be formulated as follows: are the translation-based senses and semantic relations in the Mirrors method linguistically motivated *from a monolingual point of view*? To this end, this thesis proposes to use monolingual task of WSD as a practical framework to evaluate the usefulness of the Mirrors method as a lexical knowledge source. This is motivated by the idea that a well-defined end-user application may provide a stable framework within which the benefits and drawbacks of a resource or a system can be demonstrated (e.g. Ng & Lee, 1996; Stevenson & Wilks, 2001; Yarowsky & Florian, 2002; Specia et al., 2009).

The innovative aspect of applying the Mirrors method for WSD is two-fold: first, the Mirrors method is used to obtain sense-tagged data automatically (using cross-lingual data), providing a SemCor-like corpus which allows us to exploit semantically analysed context features in a subsequent WSD classifier. Second, we will test whether training on *semantically analysed* context features, based on information from the Mirrors method, means that the system resolves other instances than a 'traditional' classifier trained on words.

In the absence of existing data sets for WSD for Norwegian, an automatically sense-tagged parallel corpus and a manually verified lexical sample of fifteen target words was developed for Norwegian as part of this thesis. The proposed automatic sense-tagging method is based on the Mirrors sense inventory and on the translational correspondents of each word occurrence. The sense-tagger provides a partially semantically analysed context—partially, because the translation-based sense-tagger can only sense-tag tokens that were successfully word-aligned. The sense-tagged English-Norwegian Parallel Corpus (the ENPC) is comparable in size to the existing SemCor.

The sense-tagged material formed the basis for a series of controlled experiments, in which the knowledge source is varied but where we maintain the same experimental framework in terms of the *classification algorithm, data sets, lexical sample and sense inventory*. First, a WSD classifier is trained on the actually co-occurring context WORDS. This knowledge source functions as a point of reference to indicate how well a traditional word-based classifier could be expected to perform, given our specific data sample and using the Mirrors sense inventory. Second, two Mirrors-derived knowledge sources were tentatively implemented, both of which attempt to generalise from the actually occurring context words as a means of alleviating the sparse data problem in WSD. For instance, if the noun *phone* was found to co-occur with the ambiguous noun *bill*N in the 'invoice' sense, and if the classifier can generalise from this to include words that are semantically close to *phone*, such as *telephone*, this means that the presence of only

one of them during learning could make both of them 'known' to the classifier at classification time.

In other words, it might be desirable to study not only word co-occurrences, as unanalysed and isolated units, but also how words enter into relations with other words (*classes of words*) in the structured network that constitutes the vocabulary of a language. In ML terms, it might be interesting to build a WSD model which learns, not how a word sense correlates with isolated words, but rather how a word sense correlates with certain classes of semantically related words. Such a tool for generalisation is clearly desirable in the face of sparse data and in view of the fact that most content words have a relatively low frequency even in larger text corpora. The first of the two Mirrors-based knowledge source rests on so-called SEMANTIC-FEATURES that are shared between word senses in the Mirrors network. Since SEMANTIC-FEATURES may include a very high number of related words, a second knowledge source was also developed—RELATED-WORDS—which attempts to selects a stricter class of near-related word senses in the wordnet-like Mirrors network.

The results indicated that the gain in abstracting from context words to classes of semantically related word senses was only marginal in that the two Mirrors-based knowledge sources only knew marginally more of the context words at classification time compared to a traditional word-based classifier. Regarding classification accuracy, the Mirrors-based SEMANTIC-FEATURES seemed to suffer from including too broad semantic information and performed significantly worse than the other two knowledge sources. The Mirrors-based RELATED-WORDS, on the other hand, was as good as, and sometimes better, than the traditional word model, but the differences were not found to be statistically significant. Although unfortunate for the purpose of enriching a traditional WSD model with Mirrors-derived information, the lack of a difference between the traditional word model and RELATED-WORDS nevertheless provides promising indications with regard to the plausibility of the Mirrors method.

# ACKNOWLEDGEMENTS

This thesis is the result of several years of research, and I wish to express my gratitude to all current and past colleagues, as well as friends and family, who have backed me up all along.

First and foremost I thank my advisor, Helge Dyvik, for his insightful and stimulating feedback that made it possible for me to complete this thesis. It is an inspiration to have a supervisor who is extremely knowledgeable and who nonetheless remains curious and open for new knowledge, thoughts and ideas. Special thanks are also due to Paul Meurer and Sindre Sørensen, without whose technical skills, knowledge of LISP and friendly patience I just might have considered to become a carpenter instead.

My PhD scholarship was awarded in affiliation to the LOGON project. The LOGON project was funded by the Norwegian Research Council's program for language technology (KUNSTI; KunnskapsUtvikling for Norsk SpråkTeknologI) and was a collaborative, Norwegian Machine Translation effort with participants from the three largest Norwegian universities (Oslo, Bergen and NTNU in Trondheim). I thank the LOGON team for providing such a supportive and inspiring environment of gifted researchers. I am in particular grateful to John A. Carroll for his interest in discussing my work and for his valuable suggestions.

After the PhD scholarship ended, I worked as a researcher at Uni Research, who also financed some months of my ph.d work. I wish to express my gratitude to the research director when I began there, Eli Hagen, and to Gisle Andersen, Knut Hofland and the rest of the 3rd floor at Uni Research.

Among my current and past colleagues at University of Bergen I must especially thank Victoria Rosén, Koenraad De Smedt, Lars G. Johnsen, Øivin Andersen and Christer Johansson for challenging discussions, theoretical and practical help along the path, and—especially Koenraad and Victoria—for hauling me over the finishing line. I also wish to highlight the value of being part of the Research School in Linguistics and Philology at the Faculty of Arts. In addition to being a social forum for me and my fellow doctorates, it also organised a so-called 'master

class', in which Nancy Ide was invited to read and discuss my work in progress. I am truly grateful to Nancy for her genuine interest in my work and for her valuable comments.

Some of the following intersect with the already mentioned categories of colleagues and some do not, in common for them is that skilled and gifted as they may be, I most of all thank them for simply being good friends: Nazareth A. Kifle, Gyri S. Losnegaard, Kjersti D. Vikøren, Anders Nøklestad and "Best" (Elisabeth Stavem). And to my family: thank you for always supporting my awkward choice of profession and for reminding me that job titles are not, at the end of the day, the most important.

Last and most importantly: Jarle, my 'bonus children' Michel, René and António, and our three-year old Victoria Helene. You have enriched my life with more fun, more energy and more laughs, and I dedicate this thesis to you.

# CONTENTS

# Part I

# Introduction

# CHAPTER 1

# THESIS INTRODUCTION

> There still remains the considerable task of identifying an "inventory-free" set of homograph-level distinctions that are useful for NLP, since they are not explicitly identified as such in any existing resource. The WSD community therefore has work to do, and should now turn itself to the task.
> (Ide & Wilks, 2006, p. 68)

## 1.1 Problem statement

How can we determine the relevant sense of an ambiguous word automatically, for instance in order to determine the appropriate translation of an ambiguous word in machine translation (MT)? This dissertation investigates the use of a translation-based lexicon resource, *the Mirrors method* (Dyvik, 2009, *inter alia*), as a knowledge source in the practical task of Word Sense Disambiguation (WSD).

Word Sense Disambiguation is the process of determining the relevant sense of an ambiguous word (henceforth: the *target word*) in context automatically. Automated WSD is motivated, not primarily 'as an end in itself', but rather as a module for higher-level systems such as machine translation (MT), information retrieval, information extraction and content analysis (see e.g. Navigli, 2009). Consider for instance an MT system that translates from Norwegian to English automatically. Given the Norwegian sentence in Example (1) below, the system needs to know which sense of the Norwegian noun *stemme* to translate; VOICE or VOTE.

(1) *Norw.* **Stemmen** hans lød plutselig interessert.

*Eng.* **??** His *vote* all of a sudden sounded interested.
*Eng.* **??** His *voice* all of a sudden sounded interested.

Although WSD is seen as an 'intermediate task' (Wilks & Stevenson, 1996), it is by no means a trivial one. The multiple meaning potential of words (and other linguistic units) poses challenges to a greater or lesser extent in most NLP tasks. Ambiguity is therefore often referred to as the major single challenge in NLP today.

In the tradition of what Abney (2000) refers to as a 're-emergence of empirical linguistics', current WSD research is dominated by a corpus-driven approach using *machine learning* (ML) techniques. The ML system 'learns' the contextual characteristics of each sense from a training corpus that contains concrete examples of contexts in which a word sense typically occurs. The learning phase usually applies statistical measures of some sort in order to analyse patterns of correlation between an ambiguous word and context words. The system may then classify previously unseen instances of the target word, based on what it has learnt from the training corpus.

### The sparse data problem

The most successful WSD approaches to date are so-called *supervised* ML approaches. Supervised learning means that since word sense information is not overtly present in raw text, each training instance is labelled with its relevant sense prior to learning. A significant obstacle facing the supervised ML methodology is the need to acquire training corpora that are

(i) sense-labelled prior to learning and

(ii) sufficiently informative for statistical methods.

Preparing sense-tagged training corpora manually is costly and time-consuming, in particular because statistical methods require large amounts of training examples: WSD as a classification problem becomes a problem of developing individual *word experts*, since the problem pertains to learning the senses of individual words. With so-called open classes (or *lexical* classes, i.e. nouns, verbs, adjectives and adverbs), the amounts of data about each word are often scarce even in larger corpus resources because lexical words have a different distribution than closed-class words (such as prepositions or determiners): closed-class words are few in number but occur often, whereas open-class words are many in number but each of them occurs comparably more rarely. Supervised machine-learning approaches to WSD therefore suffer from a sparse data problem: how can we acquire semantically annotated data on a greater scale, with minimal manual efforts? Some available sense tagged corpora exist, such as the English

SemCor (cf. Chapter (3)), but for smaller languages such corpora are rarely available. For Norwegian, in particular, there are none, and to date no substantial research has been done on WSD for Norwegian.

Traditionally, the sparse data problem is treated as a quantitative issue, implying that automated methods to acquire large amounts of data provide the solution (methods are more thoroughly discussed in Chapter (3)). This thesis pursues the idea that the size of a training corpus is only *part* of the problem: lexical content words typically have a low frequency; therefore there will inevitably be lexical gaps even in a big corpus. Consequently, a more fundamental problem of a corpus-driven approach is that the classifier remains ignorant about what *kind* of words, semantically speaking, it might expect to encounter in a test situation in general.

Consider Example (2) below. From this particular training example a classifier may learn that the ambiguous Norwegian noun *stemme*N in its VOICE sense is associated with (among other words) the noun *tone* 'note'.

(2) Sarah merket den påtatte *tonen* av likeglad interesse i **stemmen**|VOICE hans.
Sarah noticed the deliberate note of careless interest in voice-the his.
Sarah noticed the deliberate note of careless interest in his voice.

Knowing that *tone* 'note' may count as an indicator of *stemme*|VOICE is only relevant information to the extent that *tone* 'note' occurs in the context of new instances of the target word. Intuitively, it it also conceivable that words that are semantically close to *tone* 'note', such as *klang* 'pitch' or *tonefall* 'tone of voice', may also co-occur with *stemme*|VOICE.

In other words, it might be desirable to study not only words, as unanalysed and isolated units, but also how words enter into relations with other words (*classes of words*) in the structured network that constitutes the vocabulary of a language. In ML terms, it could be interesting to build a WSD model which learns, not how a word sense correlates with isolated words, but rather how a word sense correlates with certain classes of semantically related words.

Such a tool for generalisation is clearly desirable in the face of sparse data and in view of the fact that most content words have a relatively low frequency even in a larger text corpora. It is for this reason that, in spite of the dominance of statistical methods based on shallow data for the last two decades in WSD, it has always been assumed that WSD would benefit from less shallow data (Leacock & Chodorow, 1998; Mihalcea, 2002b; Resnik, 1995; see a fuller discussion in Chapter (3)). With a deeper knowledge about words and their relations, the ability to recognise word senses is no longer confined to the particular words that were observed in the set of example instances of an ambiguous word in a corpus.

### Defining senses for WSD

The second fundamental problem for WSD addressed in my thesis concerns how to define senses. WSD needs a word sense inventory which is computationally tractable and where the boundaries between senses are as clear-cut as possible. Subjectively defined sense distinctions have been the norm in WSD research; the most prominent knowledge source today being the Princeton WordNet (Fellbaum, 1998).

The interesting aspect of WordNet for WSD is that it represents exactly a kind of deeper knowledge which could provide interesting possibilities for WSD. Rather than enumerating the possible senses of a word as in common dictionaries, WordNet senses are organized in a conceptual network that expresses the semantic relations between senses. Semantic relations describe the paradigmatic dimension of lexical semantics, that is, how senses are related in terms of similarity of meaning; for instance near-synonymy (*car–automobile*), hyponymy (*car* is a hyponym to—is more specific than—*vehicle*) or hyperonymy (*vehicle* is a hyperonym to—is more general than—*car*). The term 'paradigmatic dimension' was introduced by de Saussure, who suggested a distinction between the *paradigmatic* and the *syntagmatic dimension*. The syntagmatic dimension denotes combinatorial properties between linguistic elements (*red* followed by *wine*), whereas the paradigmatic dimension encapsulates how elements may be substituted by each other (the syntagms *red wine* and *white wine* imply that *red* and *white* stand in a paradigmatic relation with respect to wine).

In terms of WSD as a problem, we may say that knowledge of contextual characteristics belongs to the syntagmatic dimension, whereas senses and the relations between them is a paradigmatic issue: the syntagmatic aspect of word meaning concerns knowledge of which words typically co-occur, whereas the paradigmatic aspect pertains to knowing that a *red wine* is a specific kind of *wine*). A resource such as WordNet provides precisely a kind of deeper knowledge about classes of semantically related words, and could therefore be of great interest for WSD.

The problem in using WordNet for WSD is that first, there are still no satisfactory ways to map WordNet senses to corpus instances reliably on a larger scale (cf. Chapter (3.5.2)). Second, the sense distinctions in WordNet are quite fine-grained, often to the extent that even humans find it difficult to assign only one of the possible senses to a given instance of the target word. This fine-grainedness makes automatic WSD unnecessarily difficult (Ide & Wilks, 2006). Furthermore, although there are projects aimed at building WordNet-like resources for other languages than English (cf. the EuroWordNet (Vossen, 1998), BalkaNet, FrameNet (Charles Fillmore & Petruck, 2003) and SIMPLE (Lenci et al., 2000), it is a fact that building such resources manually is a challenging task. This motivates the investigation of other ways of acquiring similar lexical resources, at least

semi-automatically.

## 1.2 The Mirrors as a lexical knowledge source for WSD

The proposal of this dissertation is to introduce a WordNet-like knowledge base to be used in WSD, namely the *Mirrors method* (Dyvik, 2009, *inter alia*). In recent years there has been an increased interest in the use of cross-lingual information in order to derive knowledge about word senses automatically (Tufiş et al., 2004; Pianta & Bentivogli, 2003; Brown et al., 1991; Ide et al., 2002; Dyvik, 2009). Indeed, in the 2010 SEMEVAL competition (Section (3.4.2)), an own task is devoted to this idea:

> Using translations from a corpus instead of human defined (e.g. WordNet) sense labels, makes it easier to integrate WSD in multilingual applications, solves the granularity problem that might be task-dependent as well, is language-independent and can be a valid alternative for languages that lack sufficient sense-inventories and sense-tagged corpora.
> (From the description of the SEMEVAL 2010 task #3: Cross-Lingual Word Sense Disambiguation[1])

The observation is that unrelated senses of a word tend to be lexicalized differently across languages. For instance, the 'vote' sense of the Norwegian noun *stemme*N would not normally be expected to share any translations into English with the 'voice' sense of the same word. Dyvik (1998, 2009) has developed a method which exploits the translational properties of words in order to derive word senses and the semantic relations between them automatically, resulting in a lexico-semantic network similar to the Princeton WordNet (the Mirrors method is further outlined in Chapter (4)). The Mirrors method is applicable for any language pair for which a parallel corpus and word alignment is available. The Mirrors method derives word senses and relations between word senses on a translational basis, grouping word senses that directly or indirectly share translational properties in 'semantic fields'. The relatedness between word senses in a semantic field is expressed through translation-based 'semantic features'.

Semantic features may be described as a framework for representing the semantic relatedness between word senses by assigning a unique set of semantic features to each word sense and allowing for feature inheritance: the more closely

[1] Text accessed from URL: http://semeval2.fbk.eu/semeval2.php?location=tasks, on Feb. 25. 2010

related two senses are the more features they have in common, and the more specific a sense is the more features it has in comparison to a more general sense.

As a simple example, the intuitive relatedness between *tone* 'tone', *klang* 'pitch' and *tonefall* 'tone' as kinds of *lyd* 'sound' may be expressed by all four concepts sharing a semantic feature $X$, while *tone* 'tone', *klang* 'pitch' and *tonefall* 'tone' (as hyponyms to, that is, more specific senses than, *lyd* 'sound') have one or more features each that are not shared between them (Figure (1.1)).

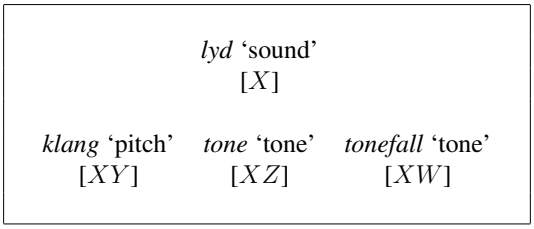



Figure 1.1: Semantic relatedness between senses, expressed through the sharing of semantic features.

Based on Mirrors information about related word senses, the observed correlation between *stemme*|VOICE and *tone* 'tone' in Example (2) above could be used to introduce concepts such as *tonefall* 'tone of voice' and *klang* 'pitch', even if they were not actually instantiated in the training material, because the two latter are found to be semantically related to *tone* 'tone' in the Mirrors word bases used in the present thesis (Example (3)).

(3) (..) den påtatte {*tone, klang, tonefall*} av likeglad interesse i **stemmen**|VOICE hans.
the deliberate {*note, pitch, tone*} of careless interest in his voice.

The appeal of the Mirrors method and its translational basis for lexical semantics is that it offers an objective and consistent—and hence, testable—criterion, as opposed to the traditional subjective judgements in lexicon classification (cf. the Princeton WordNet). But due to the lack of intersubjective "gold standards" for lexical semantics, it is not an easy task to evaluate the Mirrors method.

This thesis proposes that the monolingual task of WSD could offer a suitable evaluation framework to *evaluate* the Mirrors as a knowledge source because the basic empirical question underlying the Mirrors method is as follows: are the translation-based senses and semantic relations in the Mirrors method linguistically motivated *from a monolingual point of view*? The results reported in this dissertation are based on the language pair English–Norwegian, using an automatically word-aligned version of the English-Norwegian Parallel Corpus (ENPC).

This project sprung out of a pilot project (Lyse, 2003), in which sense-tagged data were acquired automatically based on translations from a parallel corpus and on the translation-based sense distinctions of the Mirrors method. At that time the ENPC had not been word-aligned automatically, and the project was thus limited to a small proof of concept experiment based on manually extracted translational data. Based on the promising outcome in Lyse (2003), it was desirable to test the approach on a larger scale. Such an attempt presupposed automatic word-alignment, which was under development for the ENPC while the research proposal for this thesis was being written. Being previously limited to manually extracted translational material (and having been tested on only a few words), it was not known *a priori* how well the Mirrors method would work when based on automatically word-aligned data, which would provide a larger network of translational correspondents from all automatially word-aligned tokens in a parallel corpus; additionally the quality of the automatic word-alignment was not known in advance. Therefore, a vital point when developing this project was that the WSD experiments could serve to say something about the quality of the Mirrors method when tested on a larger scale. At that time, the Mirrors method had already been evaluated against WordNet and Merriam-Webster (Thunes, 2003), finding that this kind of evaluation is problematic: when comparing manually derived (high-quality) sets of related words in WordNet against sets of related words in Merriam-Webster, the intersection even between the two high-quality resources was low.

The thought thus emerged that WSD may serve as a practical framework for evaluating the Mirrors, to a large extent inspired by interesting experiments where paradigmatic information from the Princeton WordNet is combined with corpus-based WSD (Leacock et al., 1998; Leacock & Chodorow, 1998; Montoyo & Suárez, 2001; Mihalcea, 2002b). The idea that a well-defined end-user application may provide a stable framework within which the benefits and drawbacks of a resource or a system can be demonstrated also finds support in related work (e.g. Ng & Lee, 1996; Stevenson & Wilks, 2001; Yarowsky & Florian, 2002; Specia et al., 2009).

This thesis first produces an automatically sense-tagged corpus, in which all words in running text in the ENPC that could be word-aligned are sense-tagged with Mirrors senses. The number of sense-tagged tokens in the resulting corpus exceeds that of the biggest manually tagged corpus available for English, SemCor (Section (3.4.2)). As in Example (3), we may then abstract from the actually occurring context words to Mirrors-derived semantic information about the senses of these context words. To evaluate the Mirrors-derived context information, a series of controlled experiments is established, in which the knowledge source to learn from is systematically varied while maintaining the same experimental framework in terms of the classification algorithm, data sets, lexical

sample and sense inventory. Specifically, three knowledge sources are compared:

**traditional WORD co-occurrences (WS)** This knowledge source functions as a 'best-known' point of reference to indicate how well a traditional word-based classifier could be expected to perform, given our specific data sample, sense inventory and classification algorithm.

**A SEMANTIC-FEATURE (SF) model** An automatically sense-tagged context word is replaced by the Mirrors-derived SFs associated with this word sense. This is motivated by the idea that since word senses in the Mirrors may share SFs, there could be a statistical gain in counting context words together if they are indeed related through the Mirrors method.

**RELATED-WORDS(REL-WS)** Whereas SFs may include a very high number of related words, the RELATED-WORDS definition is developed in an attempt to select a stricter class of semantically related words in the Mirrors method.

Thus, the research question may be formulated as follows:

- What is the usefulness of the Mirrors method as a lexical knowledge source for WSD?
- Can we build a WSD model using the Mirrors method which learns, not how a word sense correlates with isolated words, but rather how a word sense correlates with certain classes of semantically related words?

Under the assumption that differing senses of an ambiguous target word have different contextual correlates, we expect that if contextual evidence supports the translation-based sense distinctions of the Mirrors method, then we may take the contextual evidence to strengthen the Mirrors hypothesis. If the Mirrors turns out to be useful for WSD, this finding will be relevant to WSD research since it could alleviate the sparse data problem of WSD in two ways: first, the Mirrors method allows for an automated acquisition of sense-tagged training data, and second, it could provide enriched context information compared to a traditional corpus-based approach to WSD.

## 1.3 Thesis contributions

The principal scientific contributions of this thesis may be summed up as follows.

- This thesis is the first substantial WSD approach tested for Norwegian.

- Sense-tagged data and a lexical sample was developed (for the first time) for Norwegian. The sense-tagged English-Norwegian Parallel Corpus (the ENPC) is comparable in size to the existing SemCor. As opposed to SemCor, the sense-tagged ENPC has not been manually verified but the thesis demonstrates that it is feasible to produce large corpora on the basis of a word-aligned parallel corpus. The lexical sample contains 15 target words (ten nouns, three adjectives and two verbs) and has been manually verified. The total data set has on average 269 corpus instances, the minimum number being 54 instances and the maximum being 1324.

- An experimental knowledge source, the Mirrors method (developed by Dyvik, 1998), is explored as a purely data-driven, language-independent lexical knowledge source for WSD. The innovative aspect of applying the Mirrors method for WSD is two-fold: first, the thesis shows that the Mirrors method may be used to obtain sense-tagged data automatically (using cross-lingual data) and second, the access to a semantically anlysed, SemCor-like corpus allows us to exploit semantically analysed context features in a subsequent WSD classifier.

- The presented automatic sense-tagger requires a word-aligned parallel corpus, is language independent and sense-tags instances with perfect precision with respect to the Mirrors sense inventory.

- The thesis tests the idea that a lexical knowledge source may be evaluated within the practical settings of WSD. The Mirrors method is used as a lexical knowledge source to develop a tentative WSD model aimed at learning, not how a word sense correlates with isolated words, but rather how a word sense correlates with certain classes of semantically related words. Although the material proved to be too sparse for conclusive generalisations about the Mirrors assumptions, statistically significant differences were found between the traditional word-based classification model and the use of semantic features from the Mirrors method. This shows that the proposed experimental framework is able to point to real differences between knowledge sources.

## 1.4 Chapter overview

The rest of this dissertation is organised as follows. The current project is best described as a 'proof of concept experiment' in virtue of introducing several new resources (the Mirrors method as a lexical resource, the sense-tagged ENPC and a lexical sample for Norwegian). As opposed to WSD projects utilising well-known

resources, the present thesis therefore necessitates, first, a discussion of the theoretical foundation for this thesis in order to show why there could, at least in theory, be a gain in abstracting from traditional context words to Mirrors-derived information about the same context words; second, a careful presentation of the resources.

Part II, "Preliminaries", consists of three chapters.

**Chapter (2)** presents a reflection on the scientific legitimacy of data-driven approaches to linguistic problems.

**Chapter (3)** presents basic concepts in WSD and gives an overview of previous work.

**Chapter (4)** introduces the basic corpus resource on which this dissertation rests, the English-Norwegian Parallel Corpus (ENPC), and the Mirrors method. As will emerge from the description of the ENPC and the Mirrors method, this dissertation encountered quite a few lower-level challenges, such as the need to devise a method for selecting between lemma analyses as a module after the automatic pre-processing of the corpus.

Part III, "The Mirrors as a knowledge source for WSD" motivates and presents the experiments of this thesis. The two first chapters are particularly important in that they outline why and how the Mirrors method may be evaluated and how it may be applied in WSD experiments.

**Chapter (5)** discusses the theoretical foundation of the presented experiments by discussing different possible evaluation strategies for the Mirrors method and motivating the choice of WSD as a practical evaluation framework.

**Chapter (6)** introduces and discusses the basic knowledge sources to be compared—the traditional use of context words (WORDS) and two implementations of Mirrors-derived information about the context word (SEMANTIC-FEATURES and RELATED-WORDS). This chapter also introduces and motivates the choice to use Naive Bayes as our classification model.

**Chapter (7)** presents the automatic sense-tagging of a parallel corpus with Mirrors senses.

**Chapter (8)** motivates the choice of focussing on a lexical sample and presents a lexical sample on the basis of the sense-tagged material.

The following chapters deal with the actual experiments:

**Chapter (9)** takes a traditional WORD classifier as its starting point, testing the effect of replacing the actually occurring lemmas with information from the two Mirrors-derived knowledge sources—SEMANTIC-FEATURES and RELATED-WORDS, when available. Finally the three knowledge sources are combined in a classification setup where the most confident classifier for each test instance is allowed to vote.

**Chapter (10)** prunes away those context lemmas that were not automatically sense-tagged in context, in order to isolate the direct, theoretical effect of replacing the actually occurring words with Mirrors-derived information about the same words. A controlled experiment is also conducted to test the plausibility of the Mirrors word senses.

**Chapter (11)** provides a general conclusion and points to some future directions for research.

# Part II

# Preliminaries

# CHAPTER 2

# PERSPECTIVE: ON THE SCIENTIFIC LEGITIMACY OF DATA-DRIVEN LANGUAGE MODELLING

## 2.1 Chapter introduction

This chapter is a reflection on modern data-driven (inductive) methods in linguistics from the point of view of theory of science. It is a common perception that linguistics 'since Chomsky' has been dominated by scepticism towards the data-driven approach to linguistics, by virtue of being associated with American structuralism. The availability of large text corpora in recent years, however, has lead to what Abney (2000) refers to as a "re-emergence of empirical linguistics" (Abney, 2000; Daelemans & Bosch, 2005; Manning & Schütze, 1999, *inter alia*). In other words, it seems to be a prevailing attitude that the new interest in corpus-based studies constitutes a return to "pre-Chomskyan" ideals.

But although the renewed interest in inductive (data-driven) methods is often accompanied by references to linguistic work preceding Chomsky (notably Firth, Bloomfield and Saussure), the current interest in data-driven methods appears largely motivated by the practical access to large corpora rather than by a clear self-understanding of their relation to linguistic history and to theory of science.

An interesting question is whether the current popularity of data-driven methods means that Chomsky's criticism was in fact mistaken. More to the point,

how do data-driven methods in modern linguistics in fact represent a revival of "pre-Chomskyan" ideas?

Data-driven methods come to play in two ways in my project. First, the project intends to investigate Dyvik's hypothesis that situated translations constitute an inter-subjective and observable source of information about lexical semantics (the Mirrors method) (Dyvik, 2005, *inter alia*). Second, the Mirrors information about word senses is applied as a knowledge source in a corpus-based machine learning (ML) approach to Word Sense Disambiguation (WSD). In the Mirrors method as well as in ML methods for WSD, data are classified on the basis of evidence from a corpus.

## 2.2 'Pre-Chomskyan' ideas

> "our descriptions must be unprejudiced, if they are to give a sound basis [...]. The only useful generalizations about language are inductive generalizations." (Bloomfield, 1933) in *Language, p. 20*.

The trajectory of linguistic research known as 'structuralism' surfaced after a period of linguistic research that was dominated by the historical-comparative tradition. The comparative tradition defined language as a historical object that develops over time, and the linguist's main aim was to classify languages and to explain why they evolve as they do. De Saussure, who is often seen as the originator of linguistic structuralism, defined a division between diachronical (historical) and synchronical studies of the structural properties of a language. The synchronical perspective enabled a new view of language in which a language (and its grammar) was understood as a valid, independent object of study. This became a crucial point for American structuralism, which dominated linguistics from the 1930s through the 1960s[1].

Structuralism distanced itself from a grammatical tradition in which knowledge of specific languages (typically Latin) constituted an *a priori* model of how any other language was expected to be described. In contrast, the American structuralist Leonard Bloomfield developed an influential empiricist methodology rooted in behaviorism. Behaviorism is a psychological theory assuming that humans (and other living creatures) learn and act on the basis of observations and generalisations, which in turn is conditioned by an causal relationship between stimulus and response. Behaviourism, as a psychological theory, was useful to Bloomfield because it paved the way for a strictly antimentalist programme: only directly observable objects were valid objects of study.

[1] It should be noted that European structuralism did not necessarily have the same empiricist basis as American structuralism.

Bloomfield's strict demand for empirical observability ties his structuralism to the logic-positivistic scientific tradition. Positivism belongs to the empiristic tradition, which—as opposed to rationalism—claims that all knowledge comes from our senses. Their main enterprise is to build knowledge on that which is, or at least is conceived as being, 'positively given', that is, directly observable. They defined *verification* as a fundamental criterion for 'proper' science: for something to be scientific, there must (at least in principle) be an empirical method for direct observation. To obtain this, they recommended the method of *induction*, that is, generalisations from singular observations (in the past) to general statements (about expected future events).

## 2.3 Induction and inductivism

Let us begin by making a clear distinction between *induction* as a formal tool of logical reasoning[2], as opposed to *inductivism* as a philosophical view on science. The separation between induction and inductivism is pertinent, as we will see that induction as a method may be used without committing oneself to the inductivist view on scientific validity. Further, we will see that Chomsky's criticism mainly concerns inductivism and not induction.

The positivists, along with Bloomfield, must be characterised as inductivists. Inductivism implies that induction, a data-driven logic reasoning, is conceived as the superior method to obtain objective and scientifically valid knowledge. Inductivism fundamentally presupposes that our data (from which to induce knowledge) must be neutral in terms of theory and ethical or moral considerations.

Bloomfield emphasised the need for clear methods, or so-called "discovery procedures", in order to obtain accurate structural descriptions of a language. Without presuppositions the linguist should systematically identify the building blocks and investigate the relations between them in a bottom-up fashion. The starting point for such an analysis was a collection of language data, a *corpus*. From the corpus the linguist was to discover general rules and principles through inductive reasoning. That is to say, Bloomfield advocated a methodology in which linguistic theories and insights were not only *based* on evidence from a corpus, i.e. data-*based* (and possibly also based on other resources, such as intuition or ideas); general statements should be data-*driven*: knowledge should be extracted mechanically from a body of observable facts.

Bloomfield's data-driven classification methodology proved quite successful at the levels of phonology and morphology. With American structuralism, linguistics became acknowledged as an autonomous science, and its methodolo-

[2]Induction is a method in formal logic, along with logical deduction and abduction.

gical apparatus was carried over to new fields such as sociology and anthropology (R. Harris, 1993, p. 28).

But the research programme also had limitations and methodological problems. The most significant critic of American structuralism was Noam Chomsky, who is commonly seen as the leading exponent for a paradigm shift from the empiricist American structuralism to a rationalist research programme.

## 2.4 Criticism against American structuralism

In *Aspects of the Theory of Syntax* (1965), Chomsky criticises the way in which an emphasis on objective discovery procedures had evolved into an 'end in itself', arguing that it undermined the search for new scientific insights and posed unreasonable limits as to what the linguist could investigate. Rather than objective discoveries he advocates a 'search for insight' as a scientific guideline (*ibid.* p. 20). It is quite clear that in an approach where only observable data are valid objects of study, and in which mechanic discovery procedures on such data are seen as the only scientifically valid methodology, the emergence of new insights in language is seriously impeded.

Chomsky defined language competence, located in the brain of each individual, as his object of study. He argued that the observable speech or text data are only products of our language competence, and are thus secondary (Chomsky, 1965 p. 18). Corpora, being examples of language use, only provide indirect information about language competence. In Chomsky's opinion, the linguist should instead consult his intuition through introspection. He did not deny that a corpus may be useful as a secondary source of knowledge, but considered it *superfluous* due to intuition: why should the linguist look for examples of a sentence construction in a corpus if he knows from intuition whether the sentence is well-formed? (Chomsky, 1965 p. 4).

In his view, a corpus could not supply an independent knowledge source because it was *unreliable* and *inadequate*: Unreliable, because a corpus of language represents not only language, but also extra-linguistic features such as distractions and errors (Chomsky, 1965 p. 3); and inadequate because we cannot guarantee that a given corpus contains all linguistic phenomena.

The problem with these arguments is that we cannot guarantee that the subjective intuitions of the researcher will suffice, either, to include all possible linguistic types, nor can we guarantee that intuition offers reliable information 'untainted' by extra-linguistic factors. Humans no doubt possess intuitions about language, but as research data for the linguist, it is quite conceivable that our data and judgements of these data may become skewed towards the theoretical point that we wish to make.

(Sæbø, 2004), in his discussion of representativeness, points out that the use of corpora is crucially distinguished from introspection by virtue of enabling us to *quantify* the representativeness of linguistic phenomena. He therefore concludes that if we wish to make claims about the 'typicalness' of a phenomenon, it must be studied in a corpus (*ibid.*).

Hence, Chomsky's arguments above do not in themselves provide convincing arguments in favour of introspection, as opposed to corpus use. Chomsky rather points to a fundamental problem caused by the fact that we cannot directly access the entire language; neither through corpora nor through introspection. As a consequence he his indeed right that the inductivist objectivity ideal is problematic; but this pertains to inductivism and not to corpus use and the inductive method.

Further, Chomsky points out that since discovery procedures necessarily require directly observable features, this methodology will fail to discover linguistic phenomena that are not overtly observable, for instance syntactic ambiguity and recursion (e.g. Chomsky, 1966 p. 51). In principle we may for instance create an indefinitely long sentence by using nested relative clauses (*I saw a cat which carried a bird which..*). Chomsky perceived it as a weakness that a corpus can only generalise about what we find to represent actual language use, but it cannot state (i) what is possible in principle or (ii) what we never expect to find. Specifically, Chomsky points out that since a corpus is bound to be "finite and somewhat accidental" (Chomsky, 1957 p. 15), we may know from intuition that some things that *are* in the language just did not happen to be in our corpus. There simply is no mechanical way of distinguishing between accidental gaps in a corpus and things that we, for concrete linguistic reasons, did not expect to find. In modern empirical language modelling the view is rather that the we are interested in accounting for, not what is possible in principle, but what we actually find in observed language use (although it may not be possible to mechanically distinguish linguistic reasons for what we find from non-linguistic reasons for what we find[3]).

In the same way as Bloomfield's inductivist view on science concorded with the (at that time) dominating logical-positivism, Chomsky's criticism of American structuralism occurred in tandem with Popper's criticism of inductivism (Popper, 1959). Popper is considered as the main force behind a scientific paradigm shift from inductivism to the hypothetical-deductive method. In doing so, Popper revived Hume's criticism against induction.

According to Hume, induction is logically problematic. With deductive statements, the conclusion follows logically from its premises, such that the conclusion is less general than its premises. With induction, on the other hand, the conclusion is more general than the premises, because an inductive statement consists in generalising about an entire population based on singular observations from a

[3] I am grateful to Koenraad De Smedt for making this point.

subset. Hume states that this logically leads to an *infinite regression*. Induction is fundamentally based on an assumption that the population that we wish to say something about has uniform properties in terms of time and space: We go from singular observations to generalisations because we expect that our observations have applied, and will continue to apply, to all members of our population, also in the future. This becomes an infinite regression because we cannot from experience, i.e. inductively, verify that our observed properties always applied and will apply to those members that are not part of our material. Hence, it lies in the nature of induction that it is logically impossible to provide evidence of a general statement on the basis of singular observations. Since we cannot inductively justify a statement about uniform properties of a population, we must necessarily view this as an *a priori* true statement. But then we move away from induction, since induction is a method based on experience.

Traditional induction (as practiced by inductivists) is problematic for further reasons, too. Since induction, in the traditional inductivist view, is based on observations that are devoid of *a priori* assumptions, there is in principle no limit as to what might count as relevant data. In practice, however, we *select* data based on certain expectations about what we expect to be relevant. Concerning the inductive conclusions from a data set, we have seen that our conclusion does not follow logically from the premises. In other words, there is nothing in the method itself that assists us in choosing between conclusions. In practice, once again, we choose the conclusion that appears to be most plausible. In view of these issues, it is hard to justify that induction is a purely objective method without presuppositions, and we therefore conclude that the inductivist view on induction is not well-motivated.

Popper therefore rejected the positivist view of induction as the superior method to obtain knowledge, and also refuted their view of verification as the divisor between science and non-science. Crucially, Popper maintains that it is irrelevant how we arrived at a theory and whether we have found evidence for the theory. For instance, a theory may well have been derived using an inductive method. The important condition for a theory to be scientifically valid, to Popper, was whether it can be falsified. Since positive evidence can never 'prove' a theory, we must be able to state the conditions that would falsify our theory instead. Popper therefore recommended to replace the inductivist programme with the hypothetical-deductive method.

## 2.5 Data-driven methods in modern computational linguistics

### 2.5.1 Rule-based (knowledge-based) language modelling

Research in computational linguistics can be traced back to attempts to translate automatically between languages in the 50s. Traditionally, computational linguistics has been dominated by the rationalist approach usually associated with Chomsky (although it may be remarked that not all who use this method agree with Chomsky's specific theories). In computational linguistics, this approach is commonly termed rule-based, or knowledge-based, language modelling.

Knowledge-based modelling is characterised as a top-down approach: rather than deriving hypotheses inductively, the linguist starts from *a priori* assumptions about the underlying language system. Language modelling thus proceeds according to the hypothetical-deductive method.

### 2.5.2 Empirical language modelling

Throughout the last couple of decades there has been an increasing interest in empirical-based language modelling. It should be stressed that the general term 'empirical studies' refers to any study that uses inter-subjectively available data, manifested through linguistic corpora. By testing theory against data that are not guided by our imaginative abilities (using introspection), it is possible to discover the extent to which our theoretical assumptions concord with empirical evidence.

Empirical language modelling as a method, on the other hand, is based on the inductive idea of using singular observations to arrive at general statements. This trajectory focusses less on *a priori* assumptions about competence and emphasise the aspect of pattern recognition and experience. With the current availability of machine-readable linguistic data, it is now possible to pursue and refine some of the pre-Chomskyan ideas.

Among the researchers themselves, however, the awareness of the extent to which pre-Chomskyan research ideas are revived seems to be only tangential. Whereas the American structuralists shared Behaviorism as a theoretical framework and an inducivist view on science, the theoretical self-understanding seems less well-defined today. Instead, the primary and common driving force in current work is an interest in developing data-driven methods to categorise and predict future data.

It is not easy to understand Bloomfield's view on generalisations beyond observed data, in the way it is manifested in current practice—that is, whether research should be purely descriptive, or if the researcher should also build gener-

alising models to predict previously unseen examples of language use. On the one hand, Bloomfield states that humans can produce new utterances by analogy to previous utterances, which clearly involves a generalisation to future events. But at the same time he assumes that utterances are produced, and understood, on the basis of an inventory of lexical forms and grammatical constructions which is sufficiently limited for us to simply enumerate the forms that exist (Bloomfield, 1933, p. 37).

By contrast, the slightly later structuralist Zellig Harris (whose most famous student is Noam Chomsky) is clear in stating that the linguist should generalise beyond her observed material. His ideas have seen a resurrection in modern, corpus-based language modelling (cf. for instance Daelemans & Bosch, 2005), and we will therefore briefly sketch his view.

Z. Harris made himself a spokesman for a formal analysis of language based on statistics. He was concerned with those regularities that surface empirically through what he denominated as the *distributional relations* in language (Z. Harris, 1951, p. 5). For instance, he argued, the distributional regularities could be used to derive equivalence relations between linguistic elements (*ibid.* p. 16). Using statistical methods on what he termed a "descriptive selection of language", that is, on a corpus, it should be possible to predict relations between linguistic elements outside the collected data material, too. In other words, Z. Harris directed his attention towards the use of *empirical discoveries of regularities as a basis for establishing theoretical constructs* (for instance synonymy).

Z. Harris does not appear as a typical structuralist in the 'Bloomfieldian' sense. For instance, he does not deny that linguists use intuition and heuristic guesswork (Z. Harris, 1951, p. 1 onw.). Z. Harris argues that inductive methods may be useful as a tool to organise given observations, from which one may subsequently generalise. Since our observations, within his framework, are formally organised according to the distributional criterion, there is no possibility of what he terms an uncontrolled interpretation of data. Thus he considers methodological objectivity to be sufficiently catered for.

Zellig Harris is interesting because he recommends research which is based on inductive discovery procedures, but in his view we use the inductive method to build a system which is ultimately *deductive* (Z. Harris, 1951, p. 377 onwards). This deductive system is constituted by testable theorems, represented as predictions about the structure of possible utterances in a language. It is thus tempting to characterise Harris' proposal, not as typical of American structuralism, but as a proposal for a further development of the American structuralism. Harris' main problem, as opposed to the situation in modern linguistic research, concerns the lack of sufficient amounts of linguistic data to pursue his ideas about a statistically based linguistic analysis.

An important area of use of statistical models in linguistics today is *evaluation*;

we then use statistics to quantify how good a language model is, regardless if the model itself is statistic-based or not. This evaluation typically consists in testing the model by using it on a sample of data that represents the problem that the model is expected to solve. We will not pursue the topic of evaluation further here; our main point is that the increased interest in evaluation illustrates an important feature of modern, data-driven research: Language models and hypotheses are subject to testing and evaluation, in the spirit of Popper's hypothetical-deductive research ideal. That is to say, a revival of 'pre-Chomskyan' methodology does not counter the recognition of testability (or falsification) as the pertinent criterion of scientific activities.

We have seen that Bloomfield's structuralism is closely tied to the inductivist scientific ideal of logical positivism, through their strict demand for direct, empirical observations. What we term inductive methods in linguistics today is perhaps best understood in light of Zellig Harris' thoughts. In inductivism, a prerequisite was that data must be pure and without prior interpretation, which became one of the central criticisms against their research programme (cf. for instance Popper's statement that observations are always coloured by *a priori* assumptions). In modern linguistics the researchers rarely, if ever, claim that the selection of data proceeds without *a priori* considerations. Rather, we select those data that we, for various reasons, believe to serve a purpose in order to arrive at a meaningful language model (whether they in fact prove to be useful is an empirical question).

Furthermore, our linguistic data are often in some way analysed before we use them. Language use, or what Chomsky refers to as performance data, are in themselves understood as unanalysed linguistic data, where it is the linguist's task to analyse them. For instance, Dyvik's Mirrors method uses translations at word level in order to derive knowledge about word senses. But a parallel corpus, that is, a corpus with original texts and their translation into a target language, does not explicitly disclose exactly how words are linked to each other in terms of translations. Hence it is the researcher's job to decide *a priori* which criteria to use in order to determine when a word is said to be a proper translation of another word. Chomsky's criticism that performance data are not 'pure' thus dissolves in the new use of data-driven methods; to the contrary it is seen as essential that we must make qualified decisions about our data material in advance.

We find a similar picture in connection with machine learning (ML) approaches to WSD: For one thing, a so-called unsupervised ML classifier may induce which senses a word has based on examples of language use. But the most successful method is so-called supervised learning, in which the ML system needs training material where each instance of the ambiguous word is labelled beforehand with its appropriate sense.

Another aspect of the recent optimism evolving around a stronger focus on empirical data may be illustrated through the data-driven development of lexico-

semantic terms that we see in the Mirrors method. When looking up a word in different dictionaries, it quickly becomes clear that different lexical resources define different sense divisions for the same word. In order to develop a better understanding of the evasive status of lexico-semantic terms, intuition is of limited use, because they do not constitute a well-defined basis to begin from. It therefore seems well-motivated to explore how far it brings us to approach word senses using consistent, and hence, in principle testable, criteria, for defining sense distinctions and semantic relations between word senses.

## 2.6 Conclusion

It seems clear that methods such as the Mirrors method and supervised WSD methods are not inductive in the traditional, inductivist understanding of the term. Inductive reasoning as a method does not in itself require that the data from which to generalise must be objective. The inductivist view of the positivists required this, because induction to them represented the only and self-sufficient method to obtain scientific knowledge. The outspoken focus on evaluation that we see today, however, clearly indicates that the resulting knowledge from induction (or any other method) is not seen as 'final' knowledge, but as *hypotheses*, in line with hypotheses within the hypothetical-deductive method. Induction in modern linguistics is perhaps best seen as a methodological tool for deriving hypothetical generalisations, as an alternative to hypotheses developed through introspection.

It must be emphasised, however, that hypotheses are not 'mechanically' derived through induction. This specification follows from the fact that we do not claim to begin from 'empirically (objectively) given' observations. On the contrary, we *select* those data that we believe to be fruitful for building illuminating language models.

# CHAPTER 3

# WSD: STATE OF THE ART

> *"Ten years ago, the 'balancing act' between symbolic and statistical methods was an exciting topic for a computational linguistics workshop; today it's an apt description of the entire field."* (Resnik, 2004, p. 2)

## 3.1 Chapter introduction

The purpose of this chapter is to show how the specific idea of using a structured lexical resource (the Mirrors method) and learning from contextual semantic features is motivated in relation to current state of the art in WSD.

As we will see, the present, standard approach of applying statistical methods on corpora, and using WordNet as the main lexical knowledge source, seems to be at a halt in terms of performance (cf. for instance Navigli, 2009). There is therefore an increasing interest in alternative solutions, added semantic information being among the suggestions (Specia et al., 2009; Izquierdo et al., 2007; Patwardhan et al., 2007; Glizzio et al., 2005; Montoyo et al., 2005; Magnini et al., 2002; Cucchiarelli & Velardi, 2002; Resnik, 1995; Miller & Charles, 1991).

This chapter will not provide a comprehensive review of WSD research in general. Instead it will focus on the specifically relevant research trajectories of WSD, in particular how to overcome the knowledge acquisition bottleneck. The interested reader may find a more thorough overview of WSD in Ide and Véronis (1998); Agirre and Edmonds (2006b) and Navigli (2009).

Section (3.2) presents a formal problem statement, in which we define WSD as a two-fold knowledge source problem: first, which word senses are well-motivated to use for automated WSD; and second, how can we make a system

know how to discriminate between senses in context? The first of these problems is addressed in Section (3.3), in which knowledge sources for word senses are discussed. The second problem is discussed in Section (3.4). We will then see that the two original, main approaches seem to be gradually replaced by 'the balancing act' of combining established approaches (Section (3.5)).

## 3.2 WSD: A formal problem statement

Formally, WSD is conveniently viewed as a modelling problem. A 'model' of some object or phenomenon is a rendering of the real object where some details may have been given priority whereas other details are omitted or simplified. Modelling some real-life phenomenon is particularly useful when the real-world object or phenomenon is too complex to be approached directly, as is often the case with linguistic phenomena such as word senses.

Lexical ambiguity was noted as a challenge for the computational treatment of language in tandem with the inception of machine translation (MT) in the 1950s (cf. for instance Weaver, 1955; Kaplan, 1950; Yngve, 1955; Bar-Hillel, 1960 and Masterson, 1967). It quickly became apparent that lexical ambiguity is an extremely complex problem because of the amounts of knowledge needed. Indeed, Bar-Hillel abandoned the field of MT because he could not see how to provide the computational knowledge needed to resolve the hardest examples of lexical ambiguity.

> "The number of facts we human beings know is, in a certain very pregnant sense, infinite." (Bar-Hillel, 1960)[1]

Stevenson (2003) observes that the history of Word Sense Disambiguation is, in many ways, a history of 'the knowledge acquisition bottleneck'. This term was coined by Gale et al. (1992) and refers to the problem of finding available and useful knowledge resources for WSD. Specifically, WSD may be viewed as a twofold knowledge source problem (Figure 3.1).

On the one hand, the system needs to know which senses to choose between. Word senses are not directly observable linguistic units in the way we may observe, for instance, the grammatical inflection of words in languages such as English and Norwegian. Therefore, there is no real consensus on how to carve up the sense distinctions of a word. On the other hand, given a sense inventory to depart from, the system is to determine which of these senses is suitable in a particular context. That is to say, we must model knowledge about how and when we use a particular word in a particular meaning.

[1] Quote taken from (Gale, Church & Yarowsky, 1992)

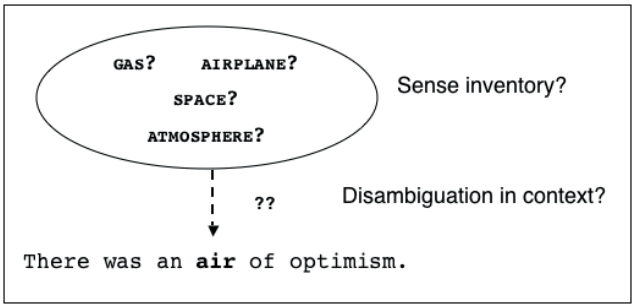



Figure 3.1: WSD as a two-fold knowledge source problem

WSD is particularly demanding because each word constitutes a separate classification task (Agirre & Edmonds, 2006a, p. 4). Other linguistic classification tasks, such as part-of-speech tagging (POS-tagging) or text categorisation, have a reasonably general, and hence limited, set of classes to model (syntactic parts of speech for POS-tagging and text types for text categorisation). For POS-tagging, for example, we only need to train *one* classifier to classify all words in a text with respect to part of speech. For WSD, by contrast, we need as many WSD classifiers as there are ambiguous words in the lexicon, since each word has a unique set of word senses. The WSD classifiers are therefore sometimes referred to as *word experts*.

When modelling lexical ambiguity as a computational problem, the relation between the model and our real-world problem may be defined as follows (Nivre, 2002): A model $M$ defines an abstract problem $Q$, which approximates a real-world problem $P$. By implication, the solution to $Q$ in a *good* model will also approximate solutions to the real-world problem $P$. Word Sense Disambiguation, then, may be approached as a classification problem, in which:

**P** = the real-world problem: determine the relevant sense of an ambiguous word in a particular context for language $L$.

**M** = the model: for instance a corpus-based probability model.

**Q** = model problem: determine the most probable sense of an ambiguous word, given a context $C$ and a sense inventory $S = \{s_1, s_2, ..., s_n\}$.

Models come about in various ways, for instance by way of manually hand-crafted representations or through a data-based approach. By data-based modelling we mean automated methods in which evidence from observed data is used to infer new information. The model is then applied on previously unseen corpus data

(test data) in *WSD classification*. A good model, then, is a model that adequately describes new (previously unseen) data.

As stated by Pedersen (1999) the primary challenge in building a good predictive model is two-fold[2]: The model should be *sufficiently complex* in the sense that it must encapsulate the important dependencies that exist, in our case, between contextual properties and each sense of an ambiguous word. On the other hand, the ability of an instantiated model to generalise to previously unseen instances of an ambiguous word hinges on *model simplicity*: if the knowledge demanded by the model is too complex, we cannot hope to acquire it consistently and reliably; and in that case we cannot make reliable predictions about new instances either.

With this in mind, it is easy to realize that a WSD model is heavily influenced by the choices we make in terms of the two modelling problems in (Figure (3.1)), namely:

- the word sense model
  A concise model for the entire vocabulary of a language in order to know which words are ambiguous and the possible senses of an ambiguous word.

- the context model
  Approaches to disambiguate words in context, by modelling the relation between context and the meaning of words.

First, the considerations for modelling word senses for WSD is discussed in Section (3.3). Then we consider approaches to disambiguate words in context in Sections 3.4 and 3.5.

## 3.3 Modelling word senses in WSD

### 3.3.1 Overview

Word senses may be approached from different points of view and for different purposes. In this thesis a translation-based view on word meaning is adopted, as developed in the Mirrors method (Dyvik, 2005). The Mirrors method and related work on word sense discovery will be discussed in Chapter (4). But Dyvik's Mirrors method is part of a theory of the epistemological basis for discovering certain semantic properties of words and is, by itself, not motivated by WSD challenges. In the remainder of this section we will therefore discuss those aspects of word senses that are specifically relevant for WSD and which fall outside the scope of Chapter (4).

[2] In Pedersen's specific discussion, he discusses WSD modelling from corpus data.

In particular, the choice of word senses for WSD is not so much driven by theoretical concerns as by their practical access and relevance for the task. Kilgariff (2006) defines 'word senses', as opposed to word *meaning*, as a lexicographic construct which represents the dictionary attempt to approximate word meaning by creating sets of discrete word senses (see Chapter (4.2) on word sense discovery). Ide and Wilks (2006, p. 55) observe that a suitable sense inventory for general-purpose WSD remains an open problem. Similarly, Palmer et al. (2006, p. 100) assert that a main question for WSD is *which sense distinctions are relevant*.

WSD systems mainly obtain their knowledge of senses from two main kinds of sources that will be discussed in the following two sub-sections:

- human-defined knowledge sources (typically lexicons, thesauri and dictionaries) (Section (3.3.2))
- data-driven sense discovery (context-driven or translation-driven) (Section (3.3.3))

### 3.3.2 Human-defined lexical resources

The typical sources of lexical knowledge in WSD today are dictionaries, thesauri and lexicons. Published lexical resources have the advantage of being publicly available for others to review them and to discuss their quality as a knowledge source; the experiments of a researcher are then also more easily replicated.

The *de facto* standard in WSD research throughout the last decade has been the Princeton WordNet (Fellbaum, 1998), although, as we will see, the resource is increasingly viewed as sub-optimal for WSD. Its drawbacks are, in part, explained by the fact that WordNet was not intended to suit the needs of researchers in computational linguistics. The Princeton WordNet was intended to model a psycholinguistic hypothesis about how humans systematise concepts (Miller, 1998), in which all words that denote the same concept are assumed to be stored together.

The crucial building blocks of WordNet are thus *synsets*, sets of words that represent one underlying lexical concept and which, by implication, are assumed to be *(near-)synonymous*. Different relations link the synonym sets, the most important ones being *hyponymy* (subconcept) and *hypernymy* (super-concept), *meronymy* (the denotation of one concept being part of the denotation of another concept), *antonymy* (opposites) and *entailment* (one concept following from another).

The lexicon has been handcrafted by lexicographers from scratch and is divided into three databases: nouns, verbs and adjectives/adverbs. The latest version (WordNet 3.0) contains some 155,000 words organised into more than 117,000

synsets[3]. A crucial feature of WordNet is that nouns and verbs are organised in a *lexical inheritance system*: A noun synset may have several hyponyms (sub-concepts), but synsets normally only have one hypernym. Because hyponymy is a transitive relation (and is one-directional), this relation produces an inheritance system. The theoretical significance of a lexical inheritance system is that if concepts inherit properties from other concepts higher up in the hierarchy, then humans may save memory by storing properties at the appropriate level, in place of repeating information for each sub-concept.

Similarly to a traditional dictionary, each concept in WordNet is supplied with a definition and a few examples of how to use the sense. As a thesaurus, WordNet links words (or, more specifically, words and concepts) in terms of semantic relations. WordNet's combined function as a counterpart to both dictionaries and thesauri, along with its architecture as a searchable electronic database, has made WordNet an attractive source of lexical knowledge for WSD. This is for instance evident in the SENSEVAL/SEMEVAL competitions, as we will see in Section (3.4.2).

**Challenges in connection with human-defined lexical resources for WSD**

The primary advantage of human-defined lexical resources for WSD is that lexical resources cover large parts of the vocabulary, and thus enable a wide-coverage approach to WSD. Furthermore, dictionaries are generally available in most languages. In the spirit of Princeton WordNet, there have been several efforts to make multilingual wordnets, too, among them EuroWordNet (Vossen, 1998), SIMPLE (Lenci et al., 2000), FrameNet (Charles Fillmore & Petruck, 2003) and BalkaNet. Recently, the Global WordNet Association was established as a free, public organisation aimed at providing a platform for discussing, sharing and connecting wordnets for all languages in the world[4].

There are two principled problems with the use of human-defined lexical resources in WSD. First, manually hand-crafted resources are typically designed for human users and not for machines. Miller (1998, p. 25) points out that a dictionary definition does not so much attempt to enumerate the full properties of word senses, as to state what *separates* the meaning of one word from other hyponyms of a more general meaning. Miller uses the example of defining a particular bird, in which case the dictionary does not state all properties that a bird has, but how this particular species differs from other kinds of birds. That is to say, the usage of a dictionary presupposes human knowledge *which is not explicit in the dictionary*. In a computational setting, dictionaries are therefore suboptimal as a knowledge

[3]The figures are taken from statistics available on the WordNet website, collected on October 12. 2009 from the following webpage: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[4]URL: http://www.globalwordnet.org. The URL was last verified on April 26, 2011.

source about word senses.

Second, dictionaries do not carve up senses according to clear, 'universal' criteria. Ide and Wilks (2006, p. 3) use the Oxford University Press as an example, which produces its English dictionary in at least four sizes; moreover the senses in the shorter dictionary are not subsets of those in a longer version. In a study on mapping dictionaries, Ide and Veronis (1990) conclude that correspondences are not always one-to-one.

So why is WordNet nonetheless the *de facto* standard in mainstream WSD research? (Palmer et al., 2006, p. 100) point to the fact that WordNet has "broad coverage, it is in the public domain and much effort has gone into linking it to WordNets of other languages." But as a lexical resource for WSD its fine-grainedness of senses has been found to cause a consistent and significant drop in system performance. After SENSEVAL-3 (2004) it was concluded that "a plateau has been reached" when using WordNet (Márquez et al., 2006, p. 187). As a result, the subsequent SEMEVAL-1 (2007) introduced two alternative tasks which only differed in the level of fine-grainedness of senses. As expected, the coarser-grained sense inventory clearly outperformed the corresponding fine-grained task.

Ide and Wilks (2006, p. 64) therefore argue that for WSD one should aim for a broader kind of distinctions that can be determined reliably from context. In addition to the quantitative arguments from WSD evaluation, they also point to theoretical arguments for focussing on homographs (unrelated senses). Homographs are often used as as a basis in the literature; for instance Ide and Wilks (2006, p. 58) mention Wierzbicka's theory of basic senses and theories of extensions from a basic lexicon, as in the work of Pustejovsky. Interestingly, they outline in some detail (*ibid.*, p. 59) the work of Klein and Murphy (2002), whose psycholinguistic experiments indicate that there are cases where etymologically related senses (i.e. non-homographs) are perceived as being as distinct as homographs (for instance, *paper* in the 'newspaper' sense as opposed to the 'material' sense). Ide and Wilks thus advocate for a level of sense distinctions that roughly corresponds to homographs. But importantly, Klein and Murphy (2002) only provide evidence for the separate representation of non-homographs, but they do not attempt to predict *which* words fall into this category.

From the point of view of relevant sense distinctions for WSD, Ide and Wilks (2006, p. 67) therefore recommend a stronger focus on *how* to identify such clearly separated sense distinctions independently of pre-existing inventories, which is the topic of the following Section (3.3.3).

### 3.3.3 Data-driven sense discovery

Data-driven methods are appealing for word senses because we as humans interpret the meaning of a word in context quite effortlessly and unconsciously, even

if we cannot necessarily agree what to *call* this sense. It therefore simply seems sensible to approach word senses in a bottom-up fashion through word sense discovery, rather than approaching word senses in a top-down approach where senses are first defined *a priori*, and only then matched with corpus data. A variety of attempts have been made to find practical means to distinguish word senses for computational purposes (syntactic behaviour, semantic and pragmatic knowledge, as well as clustering methods based on various co-occurrence measures). The two most notable criteria for data-driven sense induction that have emerged are the *distributional* and the *translational* criteria.

**(I) The distributional criterion**

> "You shall know a word by the company it keeps".
> (Firth, 1957)

The distributional hypothesis is based on the assumption that word senses may be teased out by clustering corpus instances that display similar contextual properties.

Clustering is usually treated as an unsupervised approach to WSD, (Section (3.4.3)), although it is strictly speaking a word sense discrimination method rather than a direct WSD method. In the first step, contexts that contain the particular word are collected from untagged text. Contexts are then grouped together based on measures of context similarity (usually based on statistics); using for instance word co-occurrence in global context (e.g. Yarowsky, 1992) or word co-occurrence within syntactic relations (e.g. Yarowsky, 1993). It is often discussed along with WSD approaches because context clustering also results in a classification of corpus instances.

The problem of treating clustered contexts as senses is that it is difficult to characterize what each cluster represents semantically. In particular, there is no clear criterion for predicting how many clusters a word should have, hence the user must decide the number of sense distinctions *a priori*. Also, clustering will assign each corpus instance to a group (100% coverage), even uncertain cases, at the risk of doing so at the cost of precision. For further details about clustering methods, the interested reader is referred to (Pedersen, 2006; Manning & Schütze, 1999; Navigli, 2009).

**(II) The translational criterion**

> "Meaning is manifested in the relation between languages".
> (Dyvik, 1998)

The translational hypothesis is based on the notion that translations may be seen as the product of having interpreted the meaning of the source language text (e.g. Ide & Wilks, 2006; Dyvik, 2009; Resnik & Yarowsky, 1997, 1999; Gale et al., 1992; Brown et al., 1991). Using translations to induce semantic properties of words will be further elaborated in Chapter (4); suffice it to say, for now, that there is an intuitive appeal to the use use of translations to discover word senses, since translations offer an element of inter-subjectivity in that we can usually agree about a translation regardless of theoretical points of view. It is thus epistemologically advantageous if it turns out that some of the semantic properties of words may be retrieved by studying the network of translational properties that surface in situated corpus texts (Dyvik, 2005, p. 7).

For WSD in particular, cross-lingual information is seen as a very interesting source of information for sense induction. Consider the task description for Cross-Lingual WSD in the SEMEVAL-2 (2010)[5]:

> Using translations from a corpus instead of human defined (e.g. WordNet) sense labels, makes it easier to integrate WSD in multilingual applications, solves the granularity problem that might be task-dependent as well, is language-independent and can be a valid alternative for languages that lack sufficient sense-inventories and sense-tagged corpora.

As pointed out by e.g. Márquez et al. (2006, p. 200), the use of translational data has the obvious limitation that the system can only discover those senses that are translated into different words in the other language; furthermore high-quality parallel corpora that are word-aligned are still not abundantly available.

### 3.3.4 Conclusion

We have seen that the choice of word senses for WSD is not so much driven by theoretical concerns as by their availability and relevance for the task. WordNet has been the *de facto* standard, but due to its fine-grainedness it seems that the WSD community is increasingly looking for alternative sources of knowledge about word senses. As advocated by Ide and Wilks (2006), and as indicated by the SEMEVAL-2 (2010) task of cross-lingual WSD, translational data are seen as a promising trajectory to this end.

[5]URL: http://semeval2.fbk.eu/semeval2.php?location=tasks#T8. The URL was last verified on April 26, 2011.

## 3.4 The two traditional main approaches to WSD

### 3.4.1 Overview

This section outlines the traditional main approaches to the second of the two modelling problems in (Figure (3.1)), namely that of disambiguating words in context (Word Sense Disambiguation, WSD). With the exception of extremely simple, heuristics-based models (e.g. always choosing the most frequent sense), WSD approaches attempt to model the relation between the meaning of words and the *contexts* in which they occur.

Prior to the emergence of electronically available knowledge sources for WSD in the 1980s, research on lexical ambiguity was mainly confined to 'proof of concept' investigations, in which the researcher relied on manually built resources that covered only a minuscule portion of the vocabulary. Throughout the 1970s and 1980s, lexical ambiguity was mostly addressed as an intermediate task within AI-based research on natural language processing (NLP) tasks, such as information retrieval. Ide and Véronis (1998) therefore characterise it as a watershed in WSD research when more extensive lexical resources became electronically available in the 1980s, followed by digital text corpora from the subsequent decade and onwards.

This section is organised as follows. The *evaluation* of WSD approaches, especially through the SENSEVAL/SEMEVAL competitions, has influenced the research trajectories of WSD quite significantly. Therefore it is convenient to begin this section by some background knowledge about the SENSEVAL/SEMEVAL evaluation competitions and the standard evaluation measures in (Section (3.4.2)). Then, (Section (3.4.3)) outlines the basic principles of the two traditional main approaches to WSD, viz. knowledge-based and corpus-based WSD. Each of them is illustrated with a few examples, and we then discuss their benefits and drawbacks. The discussion will show that neither of the two traditional approaches are satisfactory as stand-alone approaches. This will lead us to the discussion of what Stevenson (2003) terms 'hybrid' approaches (Section (3.5)), which is also a suitable description of the approach of the current thesis.

### 3.4.2 SENSEVAL/SEMEVAL, evaluation measures and baselines

#### SENSEVAL/SEMEVAL

SENSEVAL is an effort to provide a common framework to test different systems within the same setting (the same data sets, sense inventories and evaluation measures). Provided that a developer's system is compatible in terms of language and sense inventory, any system may participate.

In order to understand the significance of SENSEVAL, it is interesting to note that in some respects, and in particular with regard to evaluation, the history of WSD is quite reminiscent of that of machine translation (MT). MT, as WSD, is characterised as an 'AI-complete' problem that presupposes extensive 'world knowledge' (encyclopaedic knowledge) to succeed. Bar-Hillel (1960) criticised early MT for being too ambitious when aspiring towards 'fully automatic, high-quality machine translation of unrestricted text' (FAHQUT). In Bar-Hillel's opinion, the semantic complexities of the task would simply not be possible without more extensive 'world knowledge' encoded into the machines. He therefore recommended the adoption of slightly less ambitious goals.

A similar picture emerges for WSD: The steady access to new resources from the 1980s and onwards (lexical and conceptual knowledge sources as well as corpora) sparked a general optimism. From the 90s, machine learning techniques were increasingly used in the field of NLP, and good results were demonstrated for several other classification tasks. Still, the 'state of the art' assessment for WSD in the late 90s by Resnik and Yarowsky (1997, 1999) clearly reveals a gap between the expected progress in WSD and the actual situation. Comparing the task of WSD (the semantic tagging of words) to the task of part-of-speech tagging (POS-tagging, the syntactic tagging of words), Resnik and Yarowsky conclude that whereas the task of POS-tagging is 'well-understood' (facilitated by a general consensus on data sets, tag inventory and the choice of models), WSD is still not well understood.

It has been said about MT that there are as many evaluation standards as there are systems, and prior to the SENSEVAL competitions the situation was similar for WSD. As with MT, it is simply not easy to compare and evaluate WSD approaches because different methods tend to use different knowledge resources, different sense inventories and different test sets. Resnik and Yarowsky (1997, 1999) therefore highlighted the need for a standardisation of test sets (in terms of deciding on a shared set of test words, adopting a common sense inventory and establishing a shared set of test instances), in order to measure the level of progress in the field of WSD.

As a result, the first SENSEVAL competition was held in 1998. SENSEVAL (renamed SEMEVAL from 2007) has been held every three years since 1998[6]:

- SENSEVAL-1 (1998), SENSEVAL-2 (2001), SENSEVAL-3 (2004):
  Evaluation exercises focussed on WSD

- SEMEVAL-1 (2007), prospective SEMEVAL-2 (2010):
  Evaluation exercises on semantic evaluation in a broader sense, including WSD and other semantic tasks relevant for NLP.

[6]URL: http://www.senseval.org/. The URL was last verified on April 26, 2011.

The standard sense inventories of these competitions has been as follows:

- SENSEVAL-1 (1998) utilised the HECTOR sense inventory[7].
- SENSEVAL-2 (2001) adopted WordNet 1.7.
- SENSEVAL-3 (2004) used WordNet 1.7.1.
- In SEMEVAL-1 (2007), two alternative lexical sample tasks were defined: a fine-grained alternative using WordNet 2.1, and a coarse-grained alternative.

The competition has seen a steady broadening of tasks. In the first competition there was only a so-called lexical sample task, which presents a set of carefully selected ambiguous words that usually only occur once per sentence. SENSEVAL-2 introduced the task of *all-words WSD*, in which the system is to disambiguate all open-class words (i.e. nouns, verbs, adjectives and adverbs) sequentially in a text. This task requires wide-coverage systems, and hence, more stringently presupposes methods that will scale up to be applicable without the need for manual labour. The two last competitions have also had tasks specifically aimed towards cross-lingual approaches, as well as tasks that comprise semantic NLP tasks in a broader sense, such as semantic role labelling, lexical substitution, word sense induction and coreference resolution.

As regards data sets, there are two large, manually sense-tagged 'all-words' corpora for English that have become standard as development and test material in the context of SENSEVAL, and which are labelled with WordNet senses (Márquez et al., 2006, p. 173). These corpora of sense-tagged text are large by WSD standards, and will be listed in the following for later reference with respect to the automated sense-tagging experiments of the current thesis:

- *the DSO corpus*: Manually tagged with WordNet 1.5 senses; 192,800 sense-tagged instances of 121 frequent nouns and 70 verbs.
- *Semcor*: Manually tagged with WordNet 1.6 senses; ca. 234,000 sense annotations (all words in 186 files and all verbs in 166 other files).

Additionally there are five collections of context examples for a selected set of words:

[7] HECTOR was a joint Oxford University Press and Digital project in the early 1990s which resulted in a dictionary and a 20-million word corpus (which also served as a pilot for the British National Corpus). See http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/resources.html

- The SENSEVAL-1 lexical sample corpus: 41 words and 8000 instances altogether[8]

- The SENSEVAL-2 lexical sample corpus: more than 12,000 instances of 73 words (max. 100–200 training instances per word; Pedersen (2006, p. 157))

- The SENSEVAL-3 lexical sample corpus: 59 words and 12,000 instances.

- The *line-hard-serve* corpus (Leacock, Towell & Voorhees, 1993) is distributed as follows:
  The *line* corpus has more than 4000 instances of the noun *line*, distributed between six WordNet senses. The least frequent sense has 349 instances, the most frequent sense has 2218 instances.
  The *hard* data consists of more than 4000 instances of the adjective *hard*, tagged with 3 wordnet senses.
  The *serve* data contain more than 4000 instances of the verb *serve* and is tagged with 4 wordnet senses.

- The *interest* corpus (Bruce & Wiebe, 1994) has 2369 instances of the noun *interest* from the ACL/DCI Treebank and is tagged with 6 LDOCE senses.

**Evaluation metrics**

This section presents the evaluation metrics that are commonly used to quantify the performance of WSD systems. The metrics are relatively simple and commonly agreed upon, and aim at computing the relation between the total number of possible classifications and the number of correct classifications made by a WSD system. The measurements thus presuppose a test set (sometimes referred to as a 'gold standard'), in which each test instance has been annotated with its desired class manually (cf. the sense-tagged corpora in Section (3.4.2), p. 38).

*Coverage* is defined as the percentage of items in the test set for which the system makes a classification attempt, i.e. this measure does not consider if the classifications are *correct*.

$$\text{Coverage} = \frac{\text{\# classifications made}}{\text{\# Total classifications to be made}} \tag{3.1}$$

[8] It may be noted that the listed figures are taken from (Márquez et al., 2006, p. 173), but for the SENSEVAL-1 corpus a different number is specified in URL: http://www.d.umn.edu/ tpederse/data.html; namely more than 12,000 instances of 35 words

*Precision* and *recall*[9] measure the ratio between the number of correct classifications and (for precision:) the number of classifications that were actually made and (for recall:) the number of classifications that should be made. They are computed as:

$$\text{Precision} = \frac{\text{\# correct classifications}}{\text{\# classifications made}} \tag{3.2}$$

$$\text{Recall} = \frac{\text{\# correct classifications}}{\text{\# Total classifications to be made}} \tag{3.3}$$

Finally, the *F1-measure* or *balanced F-score* computes the weighted harmonic mean of precision and recall:

$$\text{F-score} = \frac{\text{2 x precision x recall}}{\text{precision + recall}} \tag{3.4}$$

Using a simple example from Navigli (2009, p. 42), a system with a precision of 100% and a recall close to zero would get approximately 50% performance if one used a simple arithmetic mean (the average of Precision + Recall). In order to produce a more balanced view, the F-score metric penalizes low values of either precision or recall, and is hence a better measure when the discrepancy between precision and recall is high.

**Baselines: Lower and upper bounds**

The lower baseline is intended as a point of reference for how well a system would perform if classification was conducted with the simplest methods. There are two common lower baselines against which any system, evaluated by the above measures, is normally compared in WSD, viz. the *random baseline* (RB) and the *most frequent sense* (MFS) baseline. The former simply predicts the result if choosing randomly between the possible senses. For instance, if there are four senses to choose between for a given ambiguous word, (RB) equals $25\%$. The most frequent sense baseline (MFS) is based on a ranking of word senses, and predicts the result when simply choosing the most frequent sense independently of context.

As opposed to these baselines, the notion of upper bound has been subject to much debate in WSD. Leacock and Chodorow (1998, p. 272) state that if one or more senses are low-frequent to such an extent that they are unlikely to ever be classified correctly, they impose an upper bound on performance: if $10\%$ of all occurrences exemplify low-frequent senses, then the upper bound on the classifier's performance will be less than $90\%$.

[9] Recall in WSD literature is sometimes referred to as *accuracy*, whereas these are separate measures in machine learning and information retrieval literature (Navigli, 2009, p. 42)

Normally, however, the upper bound in WSD is associated with *inter-annotator agreement*. This is a measure of the extent to which two or more human annotators assign the same senses to the same instances. For coarse-grained sense distinctions, the estimated inter-annotator agreement is around $90\%$ (Navigli, 2009, p. 43). For fine-grained sense inventories, such as the ones found in WordNet, inter-annotator agreement is significantly lower; estimated to be between $67$–$80\%$ (*ibid.*).

### 3.4.3 Knowledge-based WSD and corpus-based WSD

Traditionally, WSD approaches are classified as knowledge-based or corpus-based, according to their main source of knowledge to recognise meanings.

#### Knowledge-based WSD

Knowledge-based WSD relies primarily on linguistic knowledge in the shape of either hand-crafted disambiguation rules or knowledge from lexical resources such as dictionaries (e.g. Lesk, 1986; White, 1988; Ide & Veronis, 1990), thesauri (e.g. Yarowsky, 1992) and lexicons (see the following paragraph).

WordNet has become a valuable lexical resource for knowledge-based approaches due to its taxonomic structure. For instance, the earliest large-scale knowledge-based approach was that of measuring *dictionary definition overlap* (e.g. Lesk, 1986; White, 1988; Ide & Veronis, 1990). In a dictionary, a word sense is typically described in terms of words that are associated with it. The gloss often also contains a few examples of how to use the sense. Words may then be disambiguated by measuring the overlap among their sense definitions. The sense definition with the highest overlap with context words is selected as the relevant sense.

However, since the gloss of a singular entry in a dictionary typically provides too little context information for broad-coverage WSD, Banerjee and Pedersen (2003) utilised WordNet to introduce a measure of *extended gloss overlap*. Rather than comparing context words solely against the WordNet glosses of these words, the glosses of *related* concepts were also included for comparison (hypernyms, meronyms, pertainyms, etc.). They report a significant increase in performance when adding gloss information from related concepts (an increase from $18.3\%$ for the original algorithm based on only the gloss of the words in question to $34.6\%$ accuracy for their extended algorithm).

WordNet has also been exploited in a variety of *semantic similarity measures* (e.g. Agirre & Rigau, 1996; Leacock & Chodorow, 1998; Mihalcea & Moldovan, 1999; Patwardhan et al., 2005; Montoyo et al., 2005; Patwardhan et al., 2007). Co-occurring words in a given discourse are usually related in meaning (for instance a

discourse about food without food-related expressions is hardly conceivable). By using the semantic network in WordNet, one may measure the semantic similarity (i.e. the network distance) between the possible senses of co-occurring words, thus identifying the common meaning that is closest in the lexical hierarchy.

The main limitation of knowledge-based WSD is that the existing lexical resources fail to provide sufficient information about the contextual characteristics of word senses. In common dictionaries and in WordNet, each word sense is typically represented only by a short gloss and perhaps a few example phrases or sentences. Because there will typically be a multiple of ambiguity among the co-occurring words in context, the context may not have strong enough clues to make a decision towards one specific sense for one specific word. A dictionary-like resource is therefore unsatisfactory as a stand-alone knowledge source. See Mihalcea (2006) for a fuller account of knowledge-based WSD.

### Corpus-based WSD

From the 90s and onwards, increasingly large collections of electronically readable text (text corpora) became available, paving the way for the paradigm commonly referred to as *corpus-based* WSD. In contrast to the knowledge-based approach, text corpora offer numerous examples of contexts in which a word may occur.

On the basis of corpus data a machine learning (ML) system may model the contextual characteristics of each word sense, usually through probabilistic modelling. An ML algorithm has two main phases, namely *learning* (training) and *classification* (testing). The learning phase is the process of *model building*, i.e. knowledge is induced from training data, usually through a statistical analysis of the corpus examples. Classification (or testing) proceeds by predicting a class for new test instances with reference to the model.

It is common to distinguish between two main kinds of machine learning, viz. eager and lazy learning. In lazy learning, all training examples are stored in memory; for this reason lazy learning is often termed *memory-based* learning. Since no examples are forgotten, even low-frequent observations may be put to use at classification time. In the eager, or greedy, learning paradigm the system abstracts from singular (individual) observations to a general model, for instance by generating rules about characteristics of each class. Usually this means that rare exceptions from the general rule are pruned away as noise or as cases that are, at any rate, uninformative about the general cases.

There is a wide range of machine learning models in corpus-based WSD, including decision lists, decision trees, memory-based learning and bayesian modelling. As these approaches have been thoroughly outlined in detail elsewhere (e.g. Ide & Véronis, 1998; Agirre & Edmonds, 2006b; Navigli, 2009), we will

not discuss them in detail here. The model used in the experiments of the current thesis, Naive Bayes, is motivated and outlined in (Chapter (6)).

The main challenge in attempting to learn contextual characteristics of word senses from corpus data is that senses are not explicit in a text corpus. Hence, senses must either be mapped from an externally defined sense inventory to each instance of a word in the corpus (*supervised learning*) or the system must induce senses itself (*unsupervised learning*).

Unsupervised machine learning methods eschew (almost) completely external information and work directly from raw unannotated data. Being purely data-driven, the methodology is also, in principle, language-independent. In WSD, unsupervised methods are typically represented by clustering methods, which were sketched in (Section (3.3.3), p. 34). It may be noted that SENSEVAL has adopted a broadened category of 'unsupervised' that includes any approach that is *not* supervised. Any method that uses external knowledge sources is seen as 'unsupervised', as long as manual sense-tagging is not a prerequisite for the success of the method, which means that even purely knowledge-based methods or cross-lingual approaches count as 'unsupervised'.

Unsupervised methods in the limited sense of clustering methods are robust and have the advantage of not needing an externally defined lexicon resource. Still, supervised WSD has proven to outperform unsupervised methods—both in the limited and in the broader sense—consistently in a number of comparisons, including in SENSEVAL (see e.g. Resnik & Yarowsky, 1997, 1999; Stevenson & Wilks, 2001; Márquez et al., 2006 and in Navigli, 2009, p. 44-51). Corpus-based methods have thus undoubtedly had an invigorating effect on WSD research. Resnik (2004, p. 6) ascribes the success of supervised methods to the fact that supervised learning algorithms offer the advantage of "two observables: a representation of the input data, and the desired output representation".

The major limitation of the supervised approach is the need for sense-labelled data. The solution has traditionally been to sense-tag corpora manually, which is extremely costly and time-consuming. To illustrate the challenge, there are more than 75,000 WordNet sense tags for all English nouns, verbs, adjectives and adverbs (Palmer et al., 2006, p. 82). Accumulating large training material for all these word senses manually is, in the words of Leacock and Chodorow (1998, p. 265), 'simply not feasible'. This seriously limits the availability of data for new languages, new domains and new text corpora.

**Conclusion**

The two major knowledge sources for automatic WSD actually complement each other: In the knowledge-based approach, lexical knowledge sources provide sense distinctions but do not offer sufficient amounts of contextual knowledge. On the

other hand, text collections in the corpus-based approach provide abundant examples of word usage, but presupposes sense-tagged corpus material to acquire high performance. Furthermore, a fundamental problem of corpus-driven WSD is that learning is confined to those context words that occur in some training corpus; but without knowing how words relate to each other, the classifier remains ignorant about what *kind* of words it might expect to encounter in a test situation in general.

So both approaches, by themselves, face the problem of too few data. For this reason there is a major knowledge acquisition bottleneck in current WSD: how can we acquire sense-tagged data which is sufficiently informative for statistical treatment, with minimal manual efforts?

In the next section (Section (3.5)) we will review attempts to overcome the knowledge acquisition bottleneck by combining corpus evidence and lexical knowledge. The sparse data problem is traditionally treated as a quantitative issue, implying that automated methods for acquiring large amounts of data provide the solution (Section (3.5.1)). This thesis pursues the idea that the size of a training corpus is only part of the problem: lexical content words typically have a low frequency; therefore there will inevitably be lexical gaps even in a big corpus. Abstracting away from individual words may therefore be a promising alternative. This idea is motivated and developed in Section (3.5.3).

## 3.5 The 'balancing act': Hybrid approaches to WSD

So-called 'hybrid approaches' combine the benefits of knowledge-based and corpus-based WSD by building models from corpus evidence, supported by linguistic knowledge (Stevenson & Wilks, 2001; Stevenson, 2003). In the introductory chapter of the 2006 'state of the art' assessment on WSD, Agirre and Edmonds (2006a, p.18) observe that there is a "recent trend to rediscover semantic interpretation and entailment" which includes WSD as one of the component technologies, and they observe that "we are seeing a shift back to knowledge-based methods, but this time coupled with corpus-based methods".

In (Section (3.5.1)) we present three approaches to obtain sense-tagged training material in an almost, or fully, unsupervised manner—bootstrapping, monosemous relatives and cross-lingual data used for automated sense-tagging[10]. In common for these approaches is that they balance between the use of corpora and predefined resources such as mono- or bilingual dictionaries, and taxonomic re-

[10]There are also some other approaches to reduce the knowledge acquisition bottleneck, such as active learning and semantic class classifiers e.g. Villarejo, Màrquez and Rigau (2005). Since they are not directly relevant for the current thesis, the interested reader is referred to Márquez et al. (2006).

sources such WordNet. Furthermore, they have all been present in systems that achieve state-of-the-art results in WSD (see Navigli (2009) for a detailed overview of the best systems in the SENSEVAL/SEMEVAL competition up to the 2007 competition).

We then consider reasons for expecting that simply adding more corpus instances automatically will not, by itself, resolve the knowledge acquisition bottleneck in WSD. We consider the idea of combining knowledge and adding new kinds of contextual knowledge in order to enlarge, as well as to enhance, the effective size and informativeness of the training corpus (Section (3.5.2)). Indeed, the main asset of the hybrid paradigm is that it attempts to enable the integration of so-called 'deeper' knowledge into shallow corpus-based evidence, potentially yielding more accurate and comprehensive systems.

### 3.5.1 Methods for automated sense-tagging

#### Bootstrapping

In a pioneering technique, Yarowsky (1995) presented a 'bootstrapping' algorithm for WSD which requires a small set of 'seeds' to train an initial supervised classifier. The seeds exemplify some of the typical contextual surroundings of each sense. Based on the initial seeds, the classifier attempts to disambiguate word instances in an untagged corpus, and only assigns a tag if it is sufficiently confident. Confidently disambiguated instances are added to the set of context examples, providing new knowledge for the classifier in an iterative procedure.

Whereas Yarowsky (1995) adopted manually defined, binary sense divisions and used manually selected seed collocations, Mihalcea (2002a) collects seed expressions automatically from (i) corpus examples in SemCor, (ii) examples from the WordNet gloss, and (iii) examples created using the *monosemous relatives* approach (see below). Mihalcea (2002a) reports that the added 100–200 examples for each word, and a strong increase in learning was documented with the increased numbers of training instances.

A general problem with bootstrapping is that the optimal number of iterations is not *a priori* known. It is also not clear how to establish a general confidence threshold that determines when a new instance may be added to the database. Also, for the method to be fully automated, the seed collocations must be retrieved from a reliable resource, which is usually taken to mean a predefined lexical resource with examples for each sense. Since the standard resource has been WordNet, the general problems of using WordNet for WSD also limits the success of the bootstrapping approach.

### Monosemous relatives

Leacock et al. (1998) avoid the need for explicit disambiguation of a training corpus by exploiting the semantic relations between concepts in WordNet, and by taking advantage of the fact that an ambiguous word may be related in meaning to concepts that are *not* ambiguous. Such relatives are called *monosemous relatives*. On the assumption that closely related words are likely to occur in similar contexts, one may collect corpus examples of the monosemous relatives and then replace the monosemous relative by the target word prior to machine learning. Leacock et al. (1998) report that when training a classifier on *monosemous relatives*, the performance was generally only 1-2% below the level of performance training on manually tagged corpora.

Mihalcea (2002b) combine and extend the approaches of monosemous relatives and bootstrapping. Seeds for a bootstrapping algorithm are assembled from (the manually sense-tagged) SemCor, WordNet definitions and from corpus examples using the monosemous relative approach. The approach achieved promising results in the SENSEVAL-2 by building a 'supervised' classifier trained on this material. The quality of these web data was found to equal that of manually tagged data.

A problem with the monosemous relatives approach is that a monosemous relative may have other collocations than the target word itself, which may cause systematic errors in the resulting classifier, and which also means that there is a risk of losing important collocational characteristics of the actual target word (Leacock et al., 1998). The method also has some applicability restrictions. The language in question must have access to a semantic taxonomy of the type in WordNet, and the approach can only be applied for those word classes that are ordered in a semantic taxonomy (i.e. only nouns and verbs, in the case of WordNet). The method is also necessarily limited by the existence of monosemous relatives for all senses of the ambiguous word, and by the number of appearances of these monosemous relatives in the corpus. Mihalcea and Moldovan (1999) found that when restricting the semantic relations to synonyms, direct hyponyms and direct hypernyms, about 64% of the words in WordNet have monosemous "relatives" in the 30 million-word corpus of the San Jose Mercury News.

### Translation-based sense-tagging of data

This dissertation introduces a translation-based method for producing a sense-tagged corpus (Chapter (7)), which subsequently allows us to replace context words by contextual semantic features. The basic idea is the same as that of translation-driven sense discovery (cf. Section (3.3.3)), namely that the translation into another language often unveils sense distinctions, implying that translations

may be used as 'sense indicators' (e.g. Dagan, 1991; Gale et al., 1992; Dagan & Itai, 1994; Brown et al., 1991; Chan & Ng, 2005).

Gale et al. (1992) acquire sense-tagged training material automatically from a parallel corpus which has been word-aligned. As a 'proof of concept' they specifically select words where sense distinctions are clearly expressed through differing translations. Every occurrence of an ambiguous word that corresponds to one of these translations in the corpus is then 'sense-tagged'.

On the whole, previous approaches to acquire sense-tagged material automatically from translational data typically require, not only translational data, but also an external resource such as WordNet (e.g. Chan & Ng, 2005; Tufiş et al., 2004; Diab & Resnik, 2002) or a bilingual dictionary in combination with a monolingual corpus (e.g. Wang & Carroll, 2005). Such approaches are relevant to discuss under the heading of hybrids; purely data-driven approaches for sense discovery (which may or may not involve automatic sense-tagging) are discussed in (Chapter (4)).

As an example of a translation-based hybrid approach to sense-tag data automatically, Diab and Resnik (2002) combine translational data from a parallel corpus and WordNet to sense-annotate both sides of a parallel corpus. Using French-English as the language pair, the words in English that correspond to the same orthographic form in French are grouped into *target sets* (e.g. the French word *catastrophe* may be found to correspond to English {*disaster, tragedy, situation*} in the parallel corpus). This procedure is based on the intuition that even though the members may be ambiguous, their grouping in the same target set favours their shared element of meaning. Thus, each target set is associated with the WordNet sense that is closest to all members of the target set. The resulting sense associated with a target set is then mapped back to the respective correspondents in the parallel corpus. A weakness is that the approach then implicitly assumes that the "source word" that yields a target set is monosemous: If the "source word" is ambiguous, the resulting target set may contain translations that are not supposed to have any shared element of meaning.

Specia et al. (2005, 2008) produce a multilingual sense-tagged corpus for WSD, using sentence-aligned parallel corpora[11], statistical information and translation dictionaries (bilingual dictionaries generated from corpus data). The suggested approach is motivated by the need for sense-tagged data, which in particular is a problem for less studies languages. Resting on the language pair Portugese–English, they focus on a lexical sample of seven highly ambiguous English verbs (*come, get, give, go look, make* and *take*) and three other verbs (*ask, live* and *tell*). The total number of words in their compiled parallel corpus is $7,606,150$ (Eng-

[11] Sometimes the sentence alignment is many-to-one, so the corresponding 'units' may in practice be larger than just one sentence.

lish) and $7,642,048$ (Portugese) (Specia et al., 2005). The sense repository of a verb is defined as the set of all the possible translations of that verb in the corpus (the average number of possible translations for the lexical sample verbs are $203$ for the seven high-frequent ambiguous verbs and $19$ for the three other verbs) (*ibid.*, p. 2).

A "sense-tag" in the proposed approach is thus *the most probable translation*, given a predefined set of possible translations of a verb, and given a particular verb instance and a corresponding sentence. Specia et al. (2005, p. 5) report a coverage of $55\%$ of all verbs ($113,802$ found translations); i.e. coverage is low, but, similarly to the current dissertation, they highlight precision as more important, since the annotated corpus is intended to train a WSD model. As for the precision of the sense tagging process, Specia et al. (2005, p. 5) randomly selected $1,500$ annotated English instances totally; $150$ per verb. They find that on average, the approach was able to identify the correct senses of 94.2% of the tagged units while performing better on the three less ambiguous nouns. Applying the state of the art statistical alignment tool Giza++ to the same data set, the Giza++ precision was measured to $58\%$ on the same data.

Specia et al. (2008) pursue the approach further, deciding to manually review all the automatic annotations in a new parallel corpus of $5,000$ sentences for the ten verbs ($500$ sentences per verb). The corpus is automatically annotated with the translation of the verb; then 80% of the corpus is used for training a WSD classifier and 20% for testing. Their own WSD approach combines corpus-based evidence and deeper linguistic knowledge, introducing an inductive logic programming technique which has previously not been used in WSD. The performance of their WSD model is compared against that of three classical WSD models (among them Naive Bayes) on the same data, finding that their proposed inductive logic approach significantly outperforms the baseline as well as all three classical WSD models (although it is not specified exactly which contextual features the three classical features were trained with). Specia et al. (2008, p. 7) remark that a caveat of these results is that the verbs are highly ambiguous (which is natural, since each individual translation counts as an individual sense) and that it would be desirable to introduce an evaluation framework that considers what they term 'synonym translations'. As we will see, the Mirrors method attempts to group translational correspondents into partitions containing 'synonym translations'.

Summing up, using cross-lingual evidence seems to be the most promising approach for acquiring training material for supervised WSD. Consider for instance the motivation of the cross-lingual WSD task of SEMEVAL-2 (2010) (see p. 35). Unfortunately, the knowledge acquisition bottleneck is present in the fact that parallel data are scarce (although this situation might change ensuing the

SEMEVAL-2010 competition, in which the use of EuroParl for cross-lingual WSD[12] may inspire a broader interest in multilingual corpora). There have been some attempts to remedy this by using a bilingual dictionary and a monolingual corpus (e.g. Wang & Carroll, 2005), which introduces a new obstacle in that the selection of possible translations is highly limited. Another solution is to use translational data from machine translation (e.g. Diab & Resnik, 2002), which may cause both a limited selection of translations as well as spurious translations. Finally, many of the mentioned approaches lack a procedure for detecting ambiguity in *both* languages (e.g. Wang & Carroll, 2005; Diab & Resnik, 2002). As will be seen in Chapter (4), the Mirrors method considers ambiguity in both languages.

**Discussion**

Methods to acquire equivalents of manually sense-tagged corpus data remain important, since they attempt to open the knowledge acquisition bottleneck in WSD that is caused by the need for sense-tagged data. Automated sense-tagging methods are especially pertinent since they may facilitate the portability of WSD classifiers to new genres, domains, corpora and languages[13]. Glizzio et al. (2005) observe that bootstrapping is the basis of state-of-the-art systems for all-words tasks, and the monosemous relatives approach has also been part of successful experiments (e.g. Mihalcea, 2002b). Translation-based approaches to automated sense-tagging are generally considered promising, since the meaning relation between the source and target text offers an implicit 'supervision'. Also, translation-based context clusters produced by Ide et al. (2002) (see Chapter (4)) have been found to be at least as reliable as those made by human annotators.

Still, it is unclear whether they will advance progress in WSD to a great extent beyond current state of the art. Some studies indicate that larger amounts of corpus material may increase performance (Mihalcea, 2002a; Wang & Carroll, 2005; Ng, 1997b; Leacock & Chodorow, 1998), but other experiments indicate the opposite (e.g. Palmer et al., 2006, Agirre & Martínez, 2000). For instance, the *line* and *interest* corpora (Section (3.4.2), p. 38) have an entire order of magnitude greater amount of training examples in comparison to the DSO and to SemCor (p. 38), but this has not resulted in significantly higher performance.

It seems to be an emerging feeling that the shortcomings of the corpus-based approach is not only a matter of too few learning instances. In particular, it is a problem that so-called symbolic features (collocations, bag of words, word forms,

[12] URL: http://semeval2.fbk.eu/semeval2.php?location=tasks#T8. The URL was last verified on April 26, 2011.

[13] Studies show that when training on one corpus, the results on an independent corpus are generally disappointing due to different sense distributions, different relevant features, and the effect of domain and genre (Márquez et al., 2006, p. 196).

lexemes; see Section (3.5.2)) which are typically used in corpus-based WSD are extremely sparse. Consider Zipf's law (Zipf, 1935), which states that in natural languages, the frequency of a word is inversely proportional to its rank in a frequency table. Thus there will be a small set of highly frequent words (typically function words such as conjunctions and prepositions), an intermediate set of medium frequent words and a large number of low-frequent words.

Daelemans and Bosch (2005, p. 5–6) argue that the lazy learning methodology which does not discard low-frequent information (cf. Section (3.4.3)) is suitable for natural language processing (NLP) tasks because languages are generally characterized by irregularities and exceptions. Cucchiarelli and Velardi (2002), however, reject this argumentation on the basis of their own studies on the effect of different feature choices in a "classic" feature-based WSD task. They find that most features that are typically used in WSD (collocations, bag of words, word forms, lexemes), are seen only 2–3 times during learning, and they conclude that:

> It is very difficult to learn anything interesting under these circumstances: it is not a matter of "not forgetting exceptions", as suggested in Daelemans et al. (1999), since almost everything is an exception, i.e. a single occurrence phenomenon!

In other words, since lexical content words are typically low-frequent, the co-occurrences that we observe in a corpus are necessarily confined, so to say, to a random snapshot of word usage. Even if we add more examples, there will inevitably still be missing words due to the Zipfian distribution of lexical content words. Consequently, an increased number of sense-tagged instances will, in isolation, never solve the problem entirely.

The emerging awareness of the shortcomings of a stand-alone corpus-based approach is, incidentally, not unique to WSD. A similar trajectory is also found in the field of Machine Translation (MT), Oepen et al. (2007) observe that:

> Like a growing number of colleagues, we question the long-term value of purely statistical (or data-driven) approaches, both practically and scientifically. (...) Assuming sufficient training material, statistical translation quality still leaves much to be desired; and probabilistic NLP experience in general suggests that one must expect 'ceiling' effects on system evolution. (Oepen et al., 2007, p. 1)

A related position is voiced by Márquez et al. (2006), who observe that the need for sense-tagged data is only part of the challenges for the continued improvement of supervised WSD systems. For instance it remains a problem that different words have different sense inventories and different characteristics. This situation means that a WSD system needs to be highly adaptable. It has therefore become increasingly popular to build WSD systems that are a mixture of

techniques, for instance by combining the complementary benefits of knowledge-based and corpus-based approaches, in order to introduce different views on word senses.

At the level of selecting machine learning models, this trend surfaces through so-called *ensemble methods*, which is one of the most popular approaches in recent supervised WSD systems, and which is seen in many of the best performing systems in SENSEVAL-2 and SENSEVAL-3 (Márquez et al., 2006, p. 204). This technique introduces multiple views of the data by training independent classifiers based on independent kinds of contextual knowledge (lexical features, grammatical features, semantic features, etc.). Intuitively, this is a reasonable approach because the combination of different classifiers ought to reduce variance and provide a more robust system. Márquez et al. (*ibid.*, p. 204) state that "the use of ensembles helps to improve results in almost all learning scenarios and it constitutes a very helpful and powerful tool for system engineering." But at the same time, the improvement obtained by the majority of combined WSD systems is only marginal. Márquez et al. therefore think that ensemble methods in themselves are not enough in order to counter the limitations of the current supervised systems, and recommend instead to pay closer attention to the *kinds* of contextual knowledge to be used (and combined) in WSD.

### 3.5.2 Context representation: State of the art

Most WSD systems use contextual features to resolve word ambiguity[14], and as such, WSD is characterised by a high-dimensional feature space. The choice of context is quite decisive, since differing selections of context for the machine learning (ML) model may result in differences in what is learnt[15].

'Context' in WSD usually takes words as the basic unit, since the words are actually observable in a text. The easiest attainable kind of context feature is thus the plain *word form* as it occurs in the text (the conjugated form). The downside is that the different conjugational forms of the same word are then registered and counted separately, which adds to the sparse data problem. The three most common kinds of abstraction from the contextual word form are the *syntactic part of speech* (POS), the *lexical lookup form* and the *lemma*. A *lemma* is constituted by a lookup form and a part of speech, forming a family of word forms that have

[14] Although in the simplest case, one may simply choose the most frequent sense without considering context at all

[15] In fact, our *a priori* choice of context also involves a choice of models: A particular classification model, such as Naive Bayes (p. 126), is best thought of as a family of models which is instantiated in different ways, depending on the context model. As a simple analogy, this corresponds to using the term *n*-gram models to refer to a family of models, of which a bigram or a trigram are two different instantiations.

the same lexical meaning and that belong to the same conjugational paradigm (in virtue of belonging to the same syntactic part of speech). This means that a 'lemma' in this context lumps together homonymous entries of the same part of speech. Thus, *right*AJ and *right*N are two distinct lexemes with the same lexical lookup form. Additionally a word may be registered by its membership of specific collocations, of semantic classes and of subject or domain codes (Agirre & Edmonds, 2006a, p.13). One may also include syntactic information such as subcategorization or argument structure, semantic information such as selectional preferences, or pragmatic information representing the role of the word within the wider discourse (Agirre & Stevenson, 2006).

**Some basic terminology on context**

Introducing some basic terminology, we broadly speak of *local* and *topical* (or *global*) context. The former typically refers to sentence-internal information whereas the latter encapsulates information about the discourse (sentence-internal as well as in the entire discourse). We also distinguish between collocations and co-occurrences.

*Collocations* are the immediately surrounding words, and they are recorded according to their position relative to the target word. The position-specific property means that if the same piece of context information happens to occur in more than one position relative to the target word, these events will be treated separately. Collocations may be extremely useful to distinguish two senses of a word because a sense often has typical function words around it which are highly indicative of a certain sense (consider for instance the 'formation' sense of English *line*N, with its typical collocation *a line of*), or because it enters into fixed (local) phrases.

*Co-occurrences*, or topical features, denote content words that typically correlate with the target word, thus indicating the topic of the discourse. As opposed to collocations, co-occurrences are not position-specific since they are only to indicate topic (regardless of their precise position relative to the target word).

In order to select words we define a *context window* of size $n$. A collocation window of $\pm 2$, for instance, means that we collect the two immediately preceding words (the $-2$ and $-1$ positions) and the two succeeding words (the $+1$ and $+2$ positions) relative to the target word (the TW). Collocations are collected as a bag of words (BOW) that comprises the $n$ nearest co-occurrences ($n$ then denotes the *context window*). In principle one might conceive of closed-class collocations that are characteristic only for one sense of the TW, but in practice it is common to focus on open-class items (nouns, verbs, adjectives and adverbs).

Agirre and Stevenson (2006, p. 233 and onwards) analyse the experimental results conducted by several researchers, and observe that it varies for each part-of-speech and even for each word which kind of knowledge that proves beneficial for disambiguation. They therefore find that it is extremely hard to generalise about contextual properties in a way which is sufficiently complex to cover all important dependencies, and which is sufficiently simple to apply across the entire vocabulary. But importantly, they find that for all parts of speech, the combination of knowledge sources generally performs better than any feature applied individually (*ibid.* p. 241).

This is not surprising, since a system with a combinatorial design is more likely to accommodate lexical variation when applied to more than one word. A few examples from the literature will illustrate this: Mihalcea (2002b) observes that whereas one noun does not benefit from a specific kind of contextual feature (such as sense-specific keywords), the correct classification of another noun may increase by 7% using the same feature. Pedersen (2002) compared the systems that participated in the Spanish and English lexical sample tasks of SENSEVAL-2. One of these systems was the combinatorial system of Montoyo and Suárez (2001). Verbs and adjectives were disambiguated in a purely corpus-based supervised approach, whereas nouns were disambiguated in a knowledge-based approach that used textual glosses and taxonomic information from the EuroWordNet. Pedersen found that the combinatorial system of Montoyo and Suárez (2001) behaved differently from the seven other systems that took part in this SENSEVAL-2 task: on the one hand, their system resolved ambiguities at which the other systems failed, but in return the system could not disambiguate many of the instances that the other systems did resolve. Pedersen concludes that if several systems are largely in agreement, then there is little benefit in combining them since they will simply reinforce each other (*combinatorial benefits* will, on the other hand, improve performance significantly if combined appropriately).

In the next subsection we will consider in more detail the emerging trend of attempting to add semantic knowledge to corpus data.

### 3.5.3 Adding semantic knowledge to corpus data

It is not a new idea to add semantic knowledge to corpus data, neither from the point of view of lexical theory nor in previous work on WSD. But semantic knowledge sources have rarely been used in supervised WSD simply because it is not easy to integrate them in a useful way; moreover the earlier attempts to rest a WSD solely on such a knowledge source have not proven very successful.

Considering theory first, Miller and Charles (1991) discuss studies (their own and previous) that show that humans are able to make stable and reliable judgements of semantic similarity between words. For instance they report on an ex-

periment (*ibid.* pp. 12—13) in which informants are asked to rate the similarity between words. Word pairs such as *car–automobile* and *gem–jewel*, respectively, are consistently judged to be semantically closer than pairs such as *rooster–voyage* and *noon–string*. Interestingly, in the word pair *journey–car*, the informants did not judge the words to be semantically similar (to mean more or less the same thing), but intuitively they still have a strong association because a car may be used to make a journey. On the part of the informants, this kind of world knowledge seemed to be manifested as a linguistic intuition by the informants judging this word pair to have an intermediate level of similarity. Linguistically, this kind of similarity is usually described as *contextual knowledge* of words (or concepts) with a tendency to co-occur.

Such a dichotomy of word associations concords with the structuralist nomenclature of the *syntagmatic* and the *paradigmatic* dimension. The syntagmatic dimension denotes combinatorial (sequential) properties between linguistic elements (*red* followed by *wine*), whereas the paradigmatic dimension encapsulates how elements may be substituted by each other (the syntagms *red wine* and *white wine* imply that *red* and *white* stand in a paradigmatic relation). These two dimensions cover two distinct notions of semantic relatedness between words. The association between *journey–car* then exemplifies the syntagmatic aspect of word meaning (knowing which words typically co-occur), whereas the intuitive similarity between *car* and *vehicle* illustrate the paradigmatic aspect (knowing that *car* is a specific kind of *vehicle*).

Miller (1998, p. 33–34) observes that by itself, WordNet only expresses 'associations based on shared features' (such as synoymy and hyponymy) but these 'are only part of the associative structure of lexical knowledge'. The other part regards which words co-occur, i.e. we then refer to the syntagmatic kind of knowledge which is *not* systematically available in WordNet, but which is found in a text corpus.

Turning to previous work, the basic idea is to generalise from particular words to classes of words that are similar in meaning. This idea was explored as early as in 1998 by Leacock and Chodorow, who use the semantic relations in WordNet to increase "the effective size of the training data" (Leacock & Chodorow, 1998, p. 265). If, during classification, a context word is not known from training, it is looked up in WordNet, and is then measured by semantic similarity against all known context words from training. In order to measure the contribution of semantic similarity, they first use an exact collocation match, which (for 200 training instances) result in a precision of 47%. When there was no match using the local exact match, the similarity measures are used. Using this methodology, they observe a small but consistent improvement between 1–3.5%.

Mihalcea (2002b) enables generalisations from corpus data by including hypernym relations from WordNet. For instance, *kitchen* and *bedroom* are hyponyms

(subconcepts) of *room*. By generalising to a hypernym pattern of the type *room door*, this pattern will match *kitchen door* as well as *bedroom door*, even if only one of them was actually attested in the corpus. The system was tested on data provided for SENSEVAL-2, and achieved very good results for both the English all-words task and the English lexical sample task. 40.3% of the classifications are based on pattern matching in Mihalcea's experiment. (In fact, several heuristics are involved, but 51.93% of the instances are classified simply by selecting the most frequent sense.)

Mihalcea and Faruque (2004) developed SENSELEARNER, which uses a small number of hand-tagged examples and which relies heavily on the WordNet taxonomy and on SemCor. Like Leacock and Chodorow (1998), they combine local patterns with the taxonomic structure in WordNet, but whereas Leacock and Chodorow only use WordNet at classification time, Mihalcea and Faruque generalise during learning. First they extract patterns of syntactic dependencies from SemCor, such as (drink$_v$, water$_n$). Then each pair is generalized with the WordNet hypernyms of the nouns and verbs involved, thus creating semantic generalisations. In the all-words WSD task of SENSEVAL-3, this system achieved a performance of 64.6%, which placed them among the best performing systems.

There are also a few experiments that exploit the hierarchical structure of WordNet, although there is no explicit generalization. The experiment of Montoyo and Suárez (2001) (Section (3.5.2), p. 54) benefited from exploiting the semantic noun hierarchy of EuroWordNet. Montoyo et al. (2005) combine a (statistical) corpus-based WSD approach and a knowledge-based approach that uses structural knowledge from the WordNet. The knowledge-based system attempts to find if a concept is common to all the senses of the nouns that form the context by searching in the WordNet hierarchy.

Glizzio et al. (2005) exploit so-called domain relations, in which a term is associated to one or more domains (for instance a *virus* may belong to the computer domain as well as to the medicine domain, and it is therefore ambiguous). A domain model is estimated by applying term clustering on a larger, untagged corpus. A cluster then represents structured information about closely related words. Glizzio et al. thus use observed facts about the behaviour of a word sense to tentatively add new possible facts about behaviour which has not been observed so far.

Cucchiarelli and Velardi (2002) replace context nouns and verbs by their WordNet synsets or hyperonyms (generalising 1, 2 3 or 4 levels up in the taxonomy, respectively). When gradually generalizing the features, the classifier generates rules with a good generalization power. For example they cite a generated rule which reads: "if a word is preceded by a verb belonging to the synset {*observe, keep, maintain*} then it belongs to the class *psychological feature*"). Note that in itself, this concrete rule is only useful if we need to disambiguate a

word with respect to another already disambiguated word (in the example, one must first know if a verb belongs to a certain synset), which is rarely the case. Cucchiarelli and Velardi argue that such rules could nonetheless be applied to two words and all their possible sense combinations to express a preference. Second, Cucchiarelli and Velardi (2002) do not attempt to perform WSD but rather to show that in principle, it should be beneficial to abstract away from individual words.

Izquierdo et al. (2007) also investigate the idea of training classifiers that recognise semantic classes rather than individual words, and conclude (p. 159) that "another option would be to incorporate more semantic information".

## 3.6 Conclusion

This chapter has introduced some basic concepts and approaches that will be referred to in the current thesis, and we have considered the current state of the art for WSD.

For one thing, supervised, corpus-based WSD outperforms other approaches, which is probably due to the fact that this approach offers what Resnik terms 'two observables'. It may be noted that technically, it is the *combination* of predefined lexical knowledge and corpus examples that constitutes a supervised WSD approach, which is the superior approach. Moreover, many of the systems that achieve state of the art results in WSD combine corpus-based evidence and predefined resources, including mono- or bilingual dictionaries as well as taxonomic resources such as WordNet. Recent approaches focus on the use of various lexical resources and corpus-based techniques in order to avoid the substantial effort required to encode linguistic knowledge.

Second, the SENSEVAL/SEMEVAL competitions have demonstrated that WSD benefits from less fine-grained sense inventories than the one which is found in WordNet. If this leads to a less widespread use of WordNet this is in many ways regrettable, because for WSD it appears clear that added corpus examples (and automatic sense-tagging procedures) will not alone widen the knowledge acquisition bottleneck. Due to the versatility of individual words and their word sense inventories, there does not appear to be one classification model that suits each WSD classification task. Rather, it appears that in order to see advances in current WSD we should investigate new combinations of knowledge and new sources of knowledge about the correlation between word senses and context. WordNet, together with corpora, could contain exactly the kind of information that might be interesting to broaden our knowledge and understanding of word senses. For this reason, less use of WordNet might alleviate the problem of too fine-grained senses, but in return we then lose a truly valuable source of knowledge about the interrelations between word senses.

As we will see in the next chapter, these observations make the translation-based Mirrors method an interesting alternative. Based on the discussions of (Ide & Wilks, 2006) and on the cross-lingual WSD task, the cross-lingual approach seems to be the most promising for discovering word senses.

The innovative aspect of applying the Mirrors method for WSD is thus two-fold: first, we may use it to obtain sense-tagged data automatically (using cross-lingual data), which will provide a SemCor-like corpus which allows us to exploit semantically analysed context features in a subsequent WSD classifier. Second, the interesting question is then whether training on semantically analysed context features, based on information from the Mirrors method, means that the system resolves other instances than the classifier trained on words.

# CHAPTER 4

# THE MIRRORS METHOD

## 4.1 Chapter introduction

This chapter presents the Mirrors method. The Mirrors method has been developed by Helge Dyvik, University of Bergen, in the research project "From Parallel Corpus to Wordnet" (2001—2004), a cooperation project between the University of Bergen and Uni Research, department Computing, in Bergen.

On the basis of translations from a parallel corpus, the method induces senses as well as semantic relations of similarity between senses. The Mirrors method may be applied for all open-class words (nouns, verbs, adjectives and adverbs) and for any language pair for which parallel corpora exists. As adverbs do not seem to generate interesting semantic relations, they are disregarded in the current thesis.

The current dissertation uses the same data material as that reported in the development of the Mirrors method, namely the English-Norwegian Parallel Corpus (ENPC, Section (4.3)). Using translational data from the ENPC, the Mirrors has generated semantic information about almost $49,000$ English and Norwegian lemmas in the ENPC; the distribution is shown in Table (4.1) (p. 60) (figures obtained from Dyvik, 2009, p. 7). The resulting lexical entries for these ENPC lemmas are available from the Mirrors online version[1].

As the Mirrors method is well-documented elsewhere (Dyvik, 2009, 2005, 2004; Priss & Old, 2005; Lyse, 2003; Thunes, 2003; Dyvik, 1998, 1997), we will focus on the basic methodological principles and consider illustrative examples. For fuller details, the interested reader is referred to (Dyvik, 2005, *inter alia*).

[1] URL: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html. The URL was last verified on April 26, 2011.

| Language | Nouns | Verbs | Adjectives |
|---|---|---|---|
| Norwegian | 21,153 | 3,043 | 4,308 |
| English | 13,344 | 2,983 | 4,003 |

Table 4.1: Automatically derived lemma entries in the ENPC-based Mirrors database

In the following, we first assess the theoretical motivation for exploring translation-based semantics and we consider related work (Section (4.2)). Since the Mirrors requires translational data from a parallel corpus as input, we then present the parallel corpus used in the current dissertation (Section (4.3)). The subsequent section outlines the Mirrors method (Section (4.4)). Finally, we address the issue of evaluation and motivate the evaluation approach taken in the current thesis in view of previous work (Section (5.1)).

## 4.2 Theoretical motivation and related work

### 4.2.1 Theoretical motivation

The motivating question behind the Mirrors method is: How can we discover and represent lexical meaning? Today, the syntactic classification of words is fairly well-understood and it is commonly based on functional, morphological and, to some extent, semantic criteria. With the *semantic* classification of words, on the other hand, the same level of consensus has not been reached. Certain lexico-semantic terms have proven to be useful in our theoretical nomenclature, such as homonymy (semantically unrelated senses of the same form, leading to the identification of two separate lexemes), polysemy (semantically related senses of the same lexeme), and semantic relations between words (e.g. synonymy, antonymy and hyponymy). But they prove difficult to use as theoretical tools for the delimitation and representation of word meaning.

As has been presented in Chapter (3.3.3) on data-driven sense discovery, two principal criteria for sense discovery may be singled out, namely the *distributional* criterion and the *translational criterion*. According to the distributional hypothesis, the contexts in which a word may occur (the *distribution* of the word) indicate its meanings. The translational hypothesis follows the assumption that when a word in one language has multiple lexical correspondents in another language, then there "must be conceptual motivation" (Ide, 1999).

Since both hypotheses share some fundamental assumptions, they will be discussed jointly in the following. As we will see in Chapter (5.1), these common assumptions are crucial to the current dissertation, which proposes to evaluate the Mirrors method against the context-driven criterion of the distributional hypo-

thesis.

**Basic assumptions for data-driven sense induction**

Any approach to lexical meaning rests on the fundamental assumption that there are discoverable regularities in how words are used, and that these regularities correlate with our perception of the meaning of words.

In general terms, 'lexical meaning' implies that linguistic signs at word level are assumed to have a relatively stable association between their form and meaning. Without this presupposition, the task of lexico-semantic descriptions, as well as the evaluation of such descriptions, would appear meaningless: If the individual words in a sentence did not have a meaning which contributes to the meaning of the sentence there would, obviously, be nothing to say about individual words as far as meaning is concerned. Furthermore, we would be compelled to assume that meaning is given solely by the context. Such a position is spurious, since we would then expect that any two words that appear in the same context should have exactly the same meaning, which is clearly not the case (cf. the minimal sentence pair 'The monkey ate the banana' and 'The monkey ate the newspaper'). Furthermore, it is problematic to account for the creative use of words (e.g. metaphors) without the assumption that there is an underlying, stable meaning which motivates calling some uses of a word 'new', or 'creative'.

The working hypothesis in a data-driven approach to lexical meaning is that if words have certain semantic properties, and if there is a correlation between the meaning of words and how they are used, it appears plausible to hypothesize that their meanings may be discerned, at least in part, by studying their patterns of use. In a monolingual corpus one may study the distributional patterns of words; in a multilingual corpus one may study the network of translational properties. In particular, observable pattern regularities may reveal information related to the *degree of similarity* (or dissimilarity) and overlap between observed patterns, from which discriminative models may be built.

The challenge with a data-driven approach to lexical semantics is that we study patterns of *word usage*, whereas word *meaning* and other semantic properties of words are not directly observable in a corpus. A data-driven discriminative model may tell us whether two uses of a word are similar or not, but they cannot tell us *which* meanings (which concepts) these uses represent. It also means that the corpus does not objectively guide our *interpretation* of any observed patterns. For instance, we do not know the relation between the contribution of context as opposed to the contribution of the individual words; moreover, there is no clear-cut distinction between linguistic knowledge as opposed to general world knowledge.

Consider Hearst (1998), who attempts to extract semantic relations from corpus data by searching for lexico-syntactic patterns of the type "X and other Y "

(which may imply that "X is a hyponym to Y "). Textual tokens are related in ways that transcend pure semantic relations between linguistic signs, among other things tokens may be interpreted as coreferential due to extra-linguistic knowledge. For instance, Hearst's method finds that 'AIDS is a disaster', which "might be considered more a metaphorical statement than a taxonomic one" (*ibid.*, p. 139). Similarly, the Mirrors method finds the English word senses *outcome*1 and *result*3 to be semantically related to *fruit*2, which would also be considered more a metaphoric association than a taxonomic one.

The fuzzy borderline between literal meaning and context-induced meaning is also apparent when considering 'general language' as opposed to 'specialised language'. In the domain of food technology, the English noun *body* has the specialised meaning of 'richness of flavour or impression of consistency given by a product'[2], such as 'a wine with a rich, full body'. Now, if we are able to attribute the right meaning of *body* in the 'wine' context above, even without a particular expertise in food technology, should we then count this as a distinct, literal sense of *body*, or should we rather say that from some more general meaning of *body*, its precise meaning is determined from the particular context? In other words, the challenge for us is to decide "what goes into the model": what counts as relevant knowledge of meaning?

Kilgariff (2006) maintains that the relevant knowledge needed to attribute meaning to word instances in a corpus stems from a number of factors, ranging from lexicalized knowledge, general linguistic knowledge (e.g. of processes of metaphor and metonymy, as well as taxonomic knowedge), pragmatic and stylistic factors. But Kilgariff concludes that "the contexts that form the substrate of our knowledge of words and their meanings cannot be dissected into lexical and world knowledge" (*ibid.*, p. 38). Therefore he concludes that "any theory which relies on a distinction between general and lexical knowledge will founder" (*ibid.*, p. 41).

The problem with such a view is that if one rejects to draw a line between lexical knowledge and extra-linguistic knowledge, one cannot meaningfully use concepts such as lexicalisation (even though Kilgariff does discuss "the process of lexicalisation" and even grants that in the case of "real world" vocabulary, as he calls it when a word form signifies a concrete entity, we may know what the word denotes independently of context (Kilgariff, 2006, p. 38). Kilgariff thus seems to be forced, in effect, to conclude that "everything goes into the model".

Dyvik (2005, 1997) concedes that it is not given exactly where to draw the line between literal meaning and creative phenomena such as the metaphoric use of language, and that it is not even clear if there exists an absolute division. But in his view a line must inevitably be drawn somewhere, and it must be drawn relative to

[2]I am grateful to Kjersti D. Vikøren at Standards Norway for this example.

something. The problem of demarcating lexical and world knowledge then largely depends on how we choose to delimit a given language as our object of study. It may well be that in one (narrower) delimitation of language, it may be perfectly plausible to conceive of 'richness of flavour or impression of consistency given by a product' as a literal sense of *body* (i.e. a separate sense that should be listed as an individual sense), whereas the identification of a more general language would motivate that the very same use of *body* is to be understood simply due to *context*. The main point is then that this choice is not a matter of 'true' or 'false', but rather that it depends on the purpose of our description.

Dyvik (2005) refers to model-theoretic semantics after Montague in order to show that it is not a radically new thought to say that the model shapes our understanding of the semantic representations. The Montagovean set-theoretic semantic model is often limited to capturing relations of inference among expressions ("John killed Bill" entails that "Bill is dead"). An important feature of such a model is that it is "constructed, not discovered through inspection of the world, and they are constructed not so much in order to look like the world, as in order to capture relations of inference among expressions in the language" (*ibid.*, p. 6). That is, the adequacy of a semantic model does not hinge on whether or not it fully describes every aspect of meaning or of the world, but rather on its ability to provide an adequate description of that selection of linguistic phenomena at which the linguist is aiming. Our theoretical apparatus is then guided by what we consider as being relevant, given language as our object of study.

Dyvik (2005) suggests that translations may provide a way to unveil semantic properties of lexical units, by attempting to weed out translational choices that are only understood with reference to the situated text. This suggestion is motivated by the element of *predictability* to word translations. We may think of linguistically predictable translations as the stable aspects (at least synchronically speaking) of word meaning, that is, those aspects of word meaning which we would characterise as relevant when speaking of a word as being a carrier of meaning (or as being 'lexicalised'). For instance, if Norwegian *hund* 'dog' is translated as *dog*, *animal* or *bastard*, these correspondences may be said to be linguistically predictable: with a difference of precision, they serve to denote properties of *hund* 'dog' that are independent of the actual discourse from which the translations were identified. Translating *hund* 'dog' as *thief*, on the other hand, may seem perfectly reasonable in a situated discourse, but this translation is not linguistically predictable because the correct reference can only be pinned down given particular information conveyed in the discourse.

Thus, linguistically predictable translations may be interesting "precisely because semantics can be seen as the theory of unimaginative language use: the kind of use (or the aspects of use, rather) that can be accounted for purely on the basis of literal meanings." (Dyvik, 1997). Crucially, in Dyvik's view, an account of

unimaginative and literal language use is a fundamental step in order to account for the creative and unpredictable kinds of language use. Linguistic predictability, conceived in this way, means that knowledge of the linguistic meaning of words entails not only knowledge of a word in isolation, but also knowledge of its semantic relations to other words (recognising for instance a relation between *dog*, *animal* and *bastard*, while acknowledging that *thief* is generally unrelated to *dog*). Following this line of reasoning, translations may offer an interesting potential for discovering semantic properties not only about, but also between, words.

The Mirrors method, then, fundamentally rests on the observation that translational data in a parallel corpus may be conceived of as the result of a process in which the translated word has been interpreted in its context. Under the hypothesis that the translational relation between languages may be viewed as a theoretical primitive, the relation may serve as the basis for deriving various semantic properties of lexemes (Dyvik, 2005).

### 4.2.2 Related work

There is a body of work that attempts to exploit cross-lingual information to discern lexico-semantic information, but actually very few of them attempt to induce senses directly from translations (Apidianaki, 2008; Ide & Wilks, 2006; Dyvik, 2005; Ide et al., 2002; Resnik & Yarowsky, 1997, 1999; Dagan & Itai, 1994; Gale et al., 1992; Dagan, 1991; Brown et al., 1991).

For instance, a wide range of attempts have been made to use translational data to disambiguate words in a corpus, as surveyed in (Chapter (3.5.1)). These attempts do not necessarily induce senses from data (using instead external lexical knowledge sources, such as WordNet), and several of them encounter problems because their methods do not cater for ambiguity in both languages (cf. the surveyed experiments of Diab & Resnik, 2002; Wang & Carroll, 2005). There has also been some work aimed at using cross-lingual information to validate sense distinctions such as those in WordNet. Ide (1999) used translational correspondents from a six-way multilingual corpus, whereas Resnik and Yarowsky (1997, 1999) attempted to use translations that were chosen by native speakers.

As for sense induction from cross-lingual data, this still seems to be a young field of research. Ide and Erjavec (2001); Ide et al. (2002) define word senses automatically using data from a six-way multilingual corpus. Given an ambiguous word (a target word), they retrieve all corpus instances and all translations of each instance across languages. Corpus instances are then clustered according to whether they were translated the same across several languages. Importantly, Ide and Erjavec (2001) find that the resulting sense distinctions largely concord with the choices of human annotators, in particular at the coarse-grained level (cf. Chapter (3.4.2)). An advantage of their approach, as with Dyvik's Mirrors

method, is that it does not depend on statistics—it only needs one observation of a particular translational relation in order to make use of it. This is fortunate since parallel corpora, as a knowledge source, typically have smaller amounts of data than monolingual corpora. The main objection to their approach is that *multilingual* aligned corpora are extremely rare.

Apidianaki (2008) induces senses from a parallel corpus by combining contextual (distributional) information and translational information. An appealing feature of her system is that, similarly to the Mirrors method, Apidianaki obtains coarser sense distinctions by clustering translations, rather than letting each translation represent one "sense indicator".

Each possible translational equivalent (EQV) of a word $x$ is associated with a so-called *SL context* (a source language context). "Context" is here defined as the surrounding, non-hapax content words, and is basically a frequency list of all context words that co-occur with $w$ when $w$ corresponds to the particular EQV. Since the SL context list is an abstraction from individual contexts, this way of counting word frequencies circumvents the sparse data problem to some extent. Translational equivalents (EQVs) are clustered on the assumption that if two EQVs have similar source language contexts, then they are likely to point to the same sense of $w$. The resulting sense clusters contain the source language context features that led to the induction of a similarity relation between translation equivalents (and if a cluster only contains one EQV, the cluster is defined by the most informative source language words associated with this EQV).

For evaluation, Apidianaki uses an independent corpus and defines the EQV of each corpus instance of $w$ as the "reference translation" ('gold standard') of this instance. A test instance is disambiguated by comparing the source language contexts of each cluster with the context of the test instance. A sense-tag (a cluster) is defined as correct if this cluster contains the reference translation; the sense-tags may thus be evaluated using the standard WSD measures of recall and precision. The baseline is the selection of the most frequent EQV for all the instances of the polysemous word. Testing on five nouns, Apidianaki finds that the precision and recall scores clearly outperform the baseline scores for all the tested nouns.

A problem with this sense induction approach is that it uses translational data from a parallel corpus as well as statistical methods. Since clustering methods, as any statistical method, is vulnerable to sparse data, this collides with the fact that parallel corpora are usually not available in the same order of magnitude as monolingual corpora.

In the following we will present the English-Norwegian Parallel Corpus (the ENPC) which is used in the current experiments.

## 4.3 The English-Norwegian Parallel Corpus (ENPC)

### 4.3.1 Overview

The English-Norwegian Parallel Corpus (ENPC) has been developed in cooperation between the University of Oslo and Uni Computing in Bergen (Johansson et al., 1999/2002)[3]. The ENPC consists of original texts and their translations (English to Norwegian and Norwegian to English) in the domains of fiction and non-fiction, distributed according to Table (4.2). As the table shows, the corpus contains 100 original and 100 translated text extracts in each language. Altogether, the material amounts to some 2.6 million words.

| | | Fiction | Non-Fiction |
|---|---|---|---|
| English | original | 30 | 20 |
| | translated | 30 | 20 |
| Norwegian | original | 30 | 20 |
| | translated | 30 | 20 |
| | | 120 | 80 |

Table 4.2: The distribution of text in the English-Norwegian Parallel Corpus (ENPC)

The parallel corpus has been pre-processed in several ways, as shown in the overview in Figure (4.1) (p. 67). The figure shows the interaction between the preprocessing of the ENPC and where the current thesis has led to changes prior to and succeeding the Mirrors method. The ENPC has been sentence-aligned using a program based on a bilingual list of 'anchor words', also exploiting proper names, numbers and words that look the same in the two texts (Hofland & Johansson, 1998)[4]. Words are syntactically analysed through lemmatisation, through which words are assigned one or more syntactic parts of speech and lexical entries (Section (4.3.2)). The corpus has been word-aligned as part of the "From Parallel Corpus to Wordnets" project (Section (4.3.3)).

As shown in the dotted square in the figure, there are two main refinements of the ENPC ensuing the current thesis. The first concerns the analysis of words in the ENPC, where the lemmatiser sometimes left several readings of a word and where we implemented some heuristic rules for pruning away unwanted analyses (described in Section (4.3.2)). These heuristic rules apply before translational correspondents are generated as input to the Mirrors method. The Mirrors method, then, generates one word base for each open word class, containing semantic information about words. The second refinement concerns the automated sense-tagging

[3]URL: http://www.hf.uio.no/ilos/tjenester/kunnskap/sprak/omc/enpc/ENPCmanual.html. The URL was last verified on April 26, 2011.

[4]URL: http://digital.uni.no/projects/closed-projects/alignment-of-sentences. The URL was last verified on April 26, 2011.

of all nouns, verbs and adjectives on both language sides in the ENPC, based on information from the Mirrors word bases and on translational correspondences in the ENPC (described in Chapter (7)).

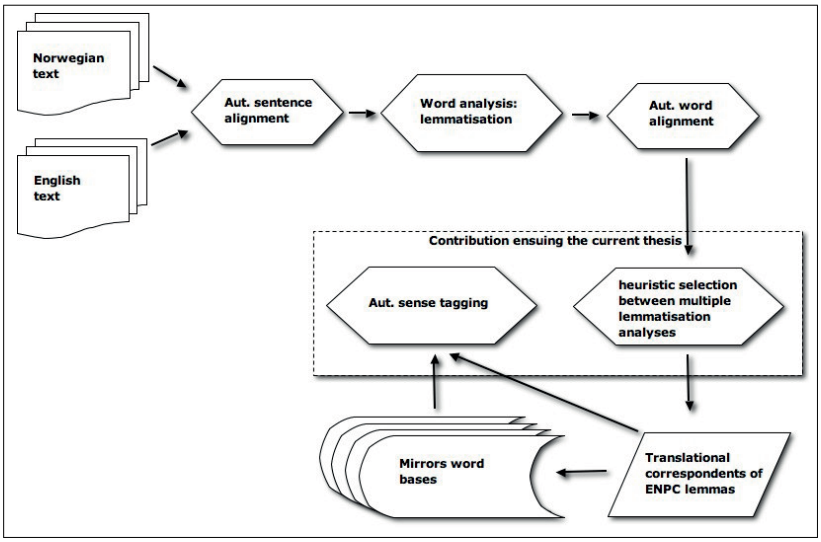



Figure 4.1: Pre-processing and linguistic analyses of the ENPC preceding and succeeding the extraction of translational input to the Mirrors method.

### 4.3.2 Lemmatisation of the ENPC: challenges and solutions

The Norwegian side of the corpus was POS-tagged and lemmatised using the Oslo-Bergen tagger[5] (henceforth: the OBT); this was done by Paul Meurer at Uni Computing. Since the OBT only has a rule set for Norwegian, the English side of the ENPC was tagged using the Penn Treebank tagset[6]. These diverging tag-sets have been unified using the EAGLES standard; this was done by Sindre Sørensen at Uni Computing, Bergen, in connection with the "From Parallel Corpus to Wordnets" project.

The OBT was evaluated in 2002 on a text sample of $30,000$ words from various text domains which included both varieties of written Norwegian (*bokmål* and

[5]URL: http://tekstlab.uio.no/obt-ny/index.html. The URL was last verified on April 26, 2011.

[6]A manual may be downloaded from URL: ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz. The URL was last verified on April 26, 2011.

*Nynorsk*)[7]. 99% of the correct tags were kept (recall) for both written varieties. Of the tags that are kept, the level of correct tags (precision) was $93,6\%$ in bokmål and $95,4\%$ in Nynorsk. This evaluation was not carried out specifically on the ENPC material.

We will devote some extra space to the OBT to motivate why and how a heuristic rule set was implemented to select among the alternative readings of a word produced by the OBT. We will begin with a few, general terminological clarifications. Linguistically, a word form (as it occurs in the corpus) belongs to an abstract *lexeme*, which is represented by the lexical entry (the lookup form) and the part of speech. The lexeme holds together a family of word forms that have the same lexical meaning and belong to the same inflectional paradigm (in virtue of belonging to the same syntactic part of speech). In other words, homonyms are linguistically defined as two distinct lexemes, where the form (spelling and/or pronunciation) is equal but the meaning differs. Since, as we will see, the Mirrors hypothesizes that linguistically motivated lexemes may be derived from translational data, the resulting senses in the Mirrors could be termed 'Mirrors-lexemes' to separate the automatically derived lexemes in the Mirrors method from the purely linguistically defined concept of a lexeme.

When producing translational correspondents as input to the Mirrors method, our point of departure is a translational relation between *lemmas* (Dyvik, 2009, p. 1). The concept of a 'lemma' is not as linguistically well-defined as a lexeme, and is for instance often found in lexicography, where the choice of lexical entry may be guided by practical considerations as well as by linguistic ones (which is why the lemmas of different lexicons may not correspond 1-to-1 to each other). In this thesis we define a lemma as a combination of a lexical lookup form and a part-of-speech, implying that a lemma holds a set of lexemes (and in the case of homonymy, the cardinality of this set will be greater than 1).

The OBT may be termed a *lemmatiser* that generates the lexical lookup forms and parts of speech that are possible given a word form, subsequently pruning away those that are not licensed given the syntactic surroundings. The lemmatiser also generates morphological information which we disregard for the purposes of the current thesis.

To illustrate why the multiple lemma alternatives are generated, Figure (4.3) (p. 69) enumerates the word forms of two Norwegian noun lemmas and one verb lemma (these examples may well be termed lexemes, since their individuation ismotivated by form as well as lexical meaning). Two of the lemmas have the same lexical entry form but have different word classes (*lage*N and *lage*V), several word forms are shared between the lemmas (*laget*, *lagene*, *lager*), and as illus-

[7]The OBT evaluation scores are available at URL: http://tekstlab.uio.no/obt-ny/english/evaluation.html. The URL was last verified on April 26, 2011.

| lemmas | word forms |
|---|---|
| **lag**N (ambiguous between 'team' and 'layer') | $lag_{sg.indef.}$<br>$laget_{sg.def.}$<br>$lag_{pl.indef.}$<br>$lagene_{pl.def.}$ |
| **lage**N (in fixed expressions: *ute av lage* 'out of order')<br>non-productive word forms still listed in a lexicon: | $lage_{sg.indef.}$<br>($laget_{sg.def.}$, $lagene_{pl.def.}$, $lager_{pl.indef.}$) |
| **lage**V ('make') | $lage_{infinitive}$<br>$lager_{present}$<br>$laga/laget/lagde_{preterite}$<br>$laga/laget/lagd_{perfectparticiple}$ |

Table 4.3: Three Norwegian lemmas illustrating shared lexical entry forms or shared word forms.

trated by the preterite and the perfect participle of the verb, some word forms have co-existing spelling variants (*laga/laget/lagde/lagd*). The second noun, *lage*N, is not a productive lexeme in modern Norwegian, but the word form *lage* exists in fixed expressions such as *ute av lage* 'out of order/balance'. Archaic uses such as *lage*N or co-existing spelling variants of the same lemma are included in the list of alternatives produced by the OBT because the OBT uses *Norsk ordbank* ('the Norwegian word bank') as a lexicon for looking up words and inflectional patterns for both written varieties of Norwegian (Bokmål and Nynorsk), but it does not know whether some alternatives are less probable than others. Since the OBT does not attempt to treat semantic ambiguity it only prunes away alternatives that are not *syntatically* licensed. Thus, word forms such as *lagene* will always have both noun alternatives (*lag*N and *lage*N) listed.

The *lagene* example illustrates a case of 'true' lexical entry ambiguity in that the two lexical entries denote distinct lexemes, even though only one of them is correct in a given context. In a second kind of cases lexical entry ambiguities arise, not because of distinct lexemes (with different meanings), but because of co-existing spelling alternatives of the same lemma. Co-existing spelling variants are not uncommon in Norwegian and are found in all open word classes (e.g. the nouns *arbeid/arbeide* 'work', *materiale/material* 'material', the verbs *gleppe/glippe* 'slip', *leite/lete* 'search' and the adjectives *fjollete/fjollet* 'silly', *uskrevet/uskreven* 'unwritten'). This second category would have been avoided if the OBT used 'normalized' variants of multiple lexical entries referring to the same lexeme (this does exist in the current reference lexicon, but they were not applied in the version of the Oslo-Bergen tagger used for this dissertation).

In both cases, the consequence is unfortunate when extracting sets of lemmas with their translational correspondents as input to the Mirrors method: the Mirrors method then records two lemmas where is really only one—in the first kind of cases, we intuitively know that only one of the readings is correct, and in the second kind of cases, several lexical entries actually point to the same lemma.

This may have unfortunate side-effects in the Mirrors method, since the Mirrors method rests on the way different lemmas overlap in terms of shared translations. Figure (4.2) (p. 70) shows the sets of translational correspondents for the lemmas *lag*N and *lage*N before implementing the heuristic rules to select between multiple lemmas. Intuitively, all the translations of *lage*N are really translational correspondents of *lag*N (although *land* and *lawyer* are erroneous word alignments that bear no relation to neither *lag*N nor *lage*N).

*lag*N:
{*air class company layer party stratum team way*}

*lage*N:
{*land layer level lawyer team*}

Figure 4.2: The sets of translational correspondents of Norwegian *lag*N and *lage*N according to the full set of lemma analyses from the Oslo-Bergen tagger

The example in Figure (4.2) shows that firstly, a lemma may happen to lose some of its translational equivalents (*layer* happens to be only associated to *lage*N and never with *lag*N). Secondly, the Mirrors method may happen to draw erroneous conclusions because of erroneous information about translational overlap among lemmas. The Mirrors assumes that if two translational correspondents of a lemma $x$ are semantically unrelated (if they point to different meanings of $x$), we do not expect them to share any other translations than $x$ itself. Ideally, for instance, we expect that *team*N and *layer*N only have Norwegian *lag*N as their shared translational correspondent. But since *team* and *layer* have *two* shared translational correspondents (*lag*N and *lage*N), they are erroneously grouped as one sense of *lag*N and of *lage*N, respectively.

Therefore, a set of heuristic rules for selecting the appropriate lemma (in cases of alternative lemmatisations with the same part of speech) was developed to be applied prior to extracting translational correspondents of lemmas as input for the Mirrors method[8]. The rule set does not interfere with the lemmatisation, but works as an extra module subsequent to the lemmatisation and the word-alignment of the parallel corpus (but before extracting the final sets of lemmas and their translational correspondents as input to the Mirrors method).

Three criteria were applied (if the first one does not apply, the second is activated; if the second does not apply we resort to the third criterion). The first criterion introduces normalised variants of lemmas that have more than one spelling alternative. The normalisation list was complied automatically by Paul Meurer for

[8]The rule set was developed through discussion between Helge Dyvik, Paul Meurer and the author, and the rule set was implemented by P. Meurer, since he has been involved in the implementation both of the Oslo-Bergen tagger and the Mirrors method.

the Norwegian and English tokens in the ENPC; then the list was manually verified by the author (in some cases a lemma in the list of spelling variants is truly ambiguous; if so we do not want to normalise it to a specific lemma).

1. *normalization*: If the lemmas listed by the Oslo-Bergen tagger belong to the same normalised lemma, they are replaced with the normalization (e.g. *arbeid, arbeide* $\rightarrow$ *arbeid*). This is also done with all unambiguous lemmatisations.

2. If unambiguous translational correspondents help to disambiguate, we use these: If a word form $x$ in L1 corresponds translationally to a word form $a$ in L2, and if $x$ has $n$ lemmatisations $\{X_1, X_2, ...X_n\}$: select $X_i$ if (1) a word form corresponding to $a$ was unambiguously analysed as an instance of $X_i$ and (2) no word forms corresponding to $a$ were unambiguously analysed as an instance of any of the alternative lemmatisations.

3. in other cases, we choose the alphabetically first lexical entry. We then risk to accidentally choose an incorrect lexical entry, but we ensure that no lexical distinctions are constructed that are not motivated by the information that is accessible.

As can be seen in the new set of translational correspondents of *lag*N and *lage*N in Figure (4.3) (p. 71), the heuristic rules work well for this particular example (and for other problematic lemmas that we observed): translations are correctly moved from *lage*N to *lag*N, while the non-productive lexeme *lage*N now only has one translational correspondent (which is actually the result of erroneous word alignment).

*lag*N:
{*air class company land lawyer layer level party stratum team way* }

*lage*N:
{*stock*}

Figure 4.3: The sets of translational correspondents of Norwegian *lag*N and *lage*N when selecting heuristically between the OBT full set of lemma analyses

### 4.3.3 Automatic word-alignment

Through the "From Parallel Corpus to Wordnets" project the ENPC has been automatically word-aligned by means of a program written by Sindre Sørensen at Uni Computing, Bergen. Word alignment was essential to extract all translations of a word in the corpus automatically. The automatic word aligner is based on a set of scoring measures, including global co-occurrence, document co-occurrence, a bilingual dictionary, sentence position, part of speech, string similarity and an anchor list. Within the Mirrors project the automatic word alignments were compared to manual alignment on a random selection of sentences across texts in the corpus. The automatic word aligner was then estimated to have a precision (correct alignment rate) of $84\%$ and an estimated recall (found alignments rate) of $62\%$.

One cannot expect all lemmas in the ENPC to be included in the Mirrors word bases, since a lemma must have been word-aligned translationally at least once to be included. This means that the number of lemmas totally in the ENPC is higher than the number of lemmas included in the Mirrors word bases. This is clearly seen in the token counts in Section (7.3.2).

## 4.4 The Mirrors method

### 4.4.1 Overview

Taking translational correspondents from a word-aligned parallel corpus as input, the Mirrors has four main stages:

- Sense discrimination
- Grouping senses into semantic fields
- Assigning semantic features to word senses
  (generating a 'lexical inheritance hierarchy')
- Deriving WordNet-like lexical entries for each word.

To illustrate each step, we first consider a Norwegian noun; since its translational correspondents are in English this example should be convenient for a non-Norwegian speaking reader. We will then consider an English noun, which (for the same reason as above) will be used as a convenient example to illustrate the derivation of semantic features and semantic relations such as synonymy and hyponymy.

A few details should be noted: Since the sense partitions of English example words contain Norwegian translational correspondents, each sense partition will be supplied with a (manually selected) English translation to the right, intended to indicate the approximate meaning of the members of the sense partition. Moreover, unless otherwise stated, all examples in this thesis are based on translational input from the automatically word-aligned version of the ENPC. Since the automatic word-alignment is not perfect, erroneous alignments may occur. Therefore, for the example words in this thesis, each alignment between an example word and its first $t$-image members has been manually verified afterwards, and all spurious first $t$-image members are marked by an asterisk (*).

### 4.4.2 Sense discrimination

In the first step, the Mirrors method takes lemmas and their sets of translational correspondents from a parallel corpus as input and individuates senses by "mirroring" translations between two languages. The Mirrors method obtains coarser sense distinctions by clustering translations, rather than assuming that each translational correspondent represents one distinct sense of the ambiguous word.

Two fundamental assumptions form the basis for the sense induction.

1. Contrastively ambigous words are not expected to have correspondents with the same ambiguity in a second language.

2. We do not expect more than one word in a language to have the same kind of ambiguity.

Consider the Norwegian noun *plan*N, for which the conventional Norwegian dictionary Bokmålsordboka[9] lists two senses. The first sense is shared between Norwegian *plan*N and English *plan*N, namely a 'scheduling' sense (e.g. plans for the future). But in addition, the Norwegian noun *plan* has a 'level' sense (e.g. levels in a building). The Norwegian noun thus represents a clear case of contrastive ambiguity, since the concept of 'planning' is semantically unrelated to 'level'. The automatic word alignment yielded the set of translational correspondences of the Norwegian noun *plan*N in (Figure (4.4)). This set of translational correspondents is referred to as the translational image (the $t$-*image*) of *plan*N. The first $t$-image is an unordered set, in which all members share some common element of meaning with the focal word of interest (our *target word*, to use the nomenclature from WSD). But crucially, not all members point back to the *same* meaning of the target word. Since we do not expect other words to share the ambiguity of *plan*N, semantically unrelated correspondences—e.g. *level* and *programme*—are only expected to share the one word *plan*N as their common translational correspondent. Semantically related words, on the other hand, such as *plan* and *programme*, may normally be expected to have more than one correspondent in common in the parallel corpus.

The next step is therefore to retrieve information about how the first $t$-image members are translated back into the source language, yielding what Dyvik (2005) terms the *inverse* $t$-*image*. (Figure (4.5)) shows that based on how the members of the first $t$-image overlap in terms of shared translations in the inverse $t$-image of *plan*N (apart from Norwegian *plan* itself), two clusters are formed in the first $t$-image of *plan*N. Specifically, English *programme, project, schedule* and *scheme*

[9]Bokmålsordboka is developed at the University of Oslo and is available online from the URL: http://www.dokpro.uio.no/ordboksoek.html. The URL was last verified on April 26, 2011.

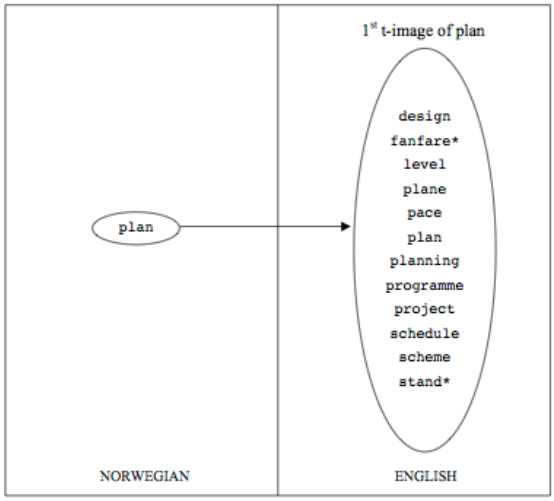



Figure 4.4: The first $t$-image of the Norwegian noun *plan*N in the ENPC. Erroneous word-alignments (manually identified *a posteriori*) are marked by an asterisk (*)

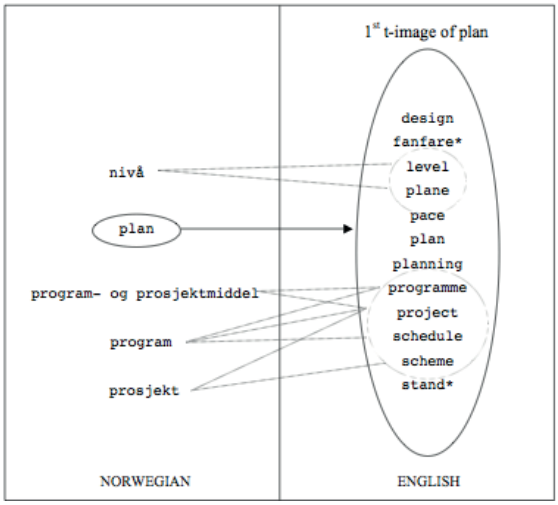



Figure 4.5: Translational overlap in the inverse $t$-image of *plan*N. The figure only includes those inverse $t$-image members that are related to more than one member of the first $t$-image.

are linked through the Norwegian correspondents of the left-hand side of the figure. Similarly, *level* and *plane* have a translational overlap through *nivå* on the Norwegian side (note, for the record, that the figure only lists those Norwegian correspondents that were relevant in virtue of uniting members of the first $t$-image; but note that the $t$-image members do have further recorded translational correspondents in their own, full $t$-images).

By this translational overlap criterion, the members of the first $t$-image of *plan*N are grouped into subsets of semantically related translations (given in Figure (4.6)). Since each subset is assumed to represent a sense of *plan*N relative to English, they are referred to as the *sense-partitions* of *plan*N's $t$-image. Henceforth, a sense partition will be referred to in terms of its numerical order in the list of senses that is shown in (4.6). For instance, the first listed sense partition, consisting of {*programme project schedule scheme*} will be referred to as Sense#1 of Norwegian *plan*N.

| **planN (Norwegian noun)** | |
|---|---|
| Sense partitions | |
| **#1:** | {*programme project schedule scheme*} |
| **#2:** | {*level plane*} |
| **#3:** | {*design*} |
| **#4:** | {*pace*} |
| **#5:** | {*plan*} |
| **#6:** | {*planning*} |
| **#7:** | {*fanfare**} |
| **#8:** | {*stand**} |

*Starred translational correspondents are false correspondents from automatic word alignment. (All correspondents in examples are manually verified *post hoc*.)

Figure 4.6: Norwegian *plan*N sense partitions

As (Figure 4.6) shows, the two first senses contain several translational correspondents of *plan*N, whereas six members of the first $t$-image of *plan*N were *not* clustered with other translational correspondents. These six correspondents form singleton sense partitions, i.e. partitions with only one member (Senses #3 through #8). Intuitively, the correspondents that were grouped together are clearly semantically related, and the sense partitions thus illustrate that the overlap criterion of the Mirrors may plausibly cluster translational correspondents into sets that approximate a 'sense'.

The example also shows that the Mirrors method displays a tendency to generate more sense-distinctions than we prefer, if the identification of contrastive ambiguity is the goal. The four sense partitions consisting of *design, pace, plan* and *planning* (Senses #3, #4, #5 and #6), respectively, could plausibly have been grouped with the other 'intention'-related word senses in Sense#1. Since there happened to be no translational correspondences that linked them with any members of Sense#1, however, the Mirrors generates five separate 'intention'-related

senses for *plan*N. In other words the Mirrors method, as most data-driven methods, is vulnerable towards sparse data. But in contrast to statistics-based approaches, the advantage of the Mirrors method is that a translational correspondence only needs to be observed once for the Mirrors method to make use of it.

Erroneous word alignment is exemplified in the two starred sense partitions (senses#7 and #8); i.e. *plan*N never corresponded with *fanfare* or *stand*. Since erroneous word alignments are typically sporadic, they tend to be also low-frequent (typically occurring just once). Therefore, Dyvik attempted earlier in the course of this dissertation to weed out all alignments that were only registered once. An analysis of the resulting Mirrors word bases indicated that one did lose a lot of noise but also quite a few important translational correspondents; therefore this attempt was subsequently abandoned.

With regard to using the Mirrors as a knowledge source in Word Sense Disambiguation, it should be noted that redundant senses and word alignment errors need not become a major source of error for WSD, because they are typically low-frequent. For instance, each of the singleton sense partitions in (Figure (4.6)) only corresponded once with *plan*N in the ENPC. In other words, only one corpus instance will be tagged with the 'design' sense#3, one will be tagged with the 'pace' sense#4, etc. Similarly, word alignment errors also tend to be sporadic; the two starred entries in (Fig. 4.6) only occurred once each. Being this low-frequent, they will ignored as senses in a WSD classification scheme, because one or two instances of a sense is not enough. Hence, unless word alignment errors devastate the Mirrors sense induction (if spurious translational overlaps are created), they need not become a dominant source of error when using this material for WSD. A far more devastating outcome for sense-tagging would be if the Mirrors tended to group unrelated senses together, which does not often seem to be the case.

### An English example: *gap*N

We will now consider the English noun *gap*N. This noun was chosen as an example of a non-homograph which intuitively still has quite clear sense divisions, namely that of 'opening, void' on the one hand and 'difference' on the other. A homograph is a word with sense divisions that are etymologically unrelated. Ide and Wilks (2006, p. 13) discuss examples of non-homographs that, according to psycho-linguistic experiments, seem as distinct to humans as homographs. This 'in turn suggests that they may be just as relevant for NLP' (*ibid.* p.13). They mention for instance the English noun *paper*N, where its semantically related senses of 'sheet of paper' and 'newspaper' are lexicalised differently in French as *journal* and *papier*, respectively. The seeming agreement between some psycho-linguistic evidence and translational data makes the cross-lingual approach to lexical semantics particularly interesting. (The Mirrors method, incidentally, generates nine

sense partitions of the English noun *paper*N based on automatic word-alignment, as can be scrutinised at the Mirrors online[10]. A manual inspection of these showed that there were quite a few erroneous correspondents, making the noun less suitable as a pedagogical example, but it did isolate the 'newspaper' sense from a larger group comprising 'piece of paper' as well as 'things written on paper').

The english noun *gap*N is listed with the following senses in the Princeton WordNet online[11] and in Merriam-Webster[12], respectively (Figure (4.4)). In order to show the level of agreement between the two resources, the listed concepts have been arranged next to each other, where appropriate. Each sense is enumerated according to its order in WordNet and Merriam-Webster, respectively. For instance, the senses (i) and (iv) in Merriam-Webster correspond pretty well to the WordNet sense (ii) and (iii) (the difference between the WordNet senses (ii) and (iii) seems to be that they denote a larger as opposed to a smaller opening, respectively). Although *gap*N does not have etymologically unrelated sense distinctions, the two lexical resources enumerate quite similar senses, which indicates that the sense divisions are indeed relatively clear.

| **Princeton WordNet** | | **Merriam-Webster** | |
|---|---|---|---|
| (i) | disparity, difference | (vii) | disparity, difference |
| (ii) | opening; open or empty space in or between things | (i) | break/gap in a barrier |
| (iii) | narrow opening,crack | (iv) | separation in space, or incomplete/deficient area (e.g. knowledge gap) |
| (iv) | a pass between mountain peaks | (ii) | mountain pass, ravine |
| (v) | difference in opinions, views or situations | (viii) | difference in character or attitude |
| | | (viiii) | a problem caused by difference (e.g. communication gap) |
| (vi) | break, interruption of continuity | (v) | break in continuity, hiatus |
| | | (iii) | spark gap |
| | | (vi) | a break in the vascular cylinder of a plant where a vascular trace departs from the central cylinder |

Table 4.4: WordNet vs. Merriam-Webster senses of the English noun *gap*N. The senses are arranged according to meaning in order to show the similarities and differences of senses defined by the two lexical resources.

If we now turn to the Mirrors method, the first $t$-image of *gap*N is given in (Figure (4.7)) and its resulting sense partitions in (Figure (4.8)). Since the non-Norwegian reader may not know the Norwegian translations, an English translation has been added in quotes to denote the contents of each sense partition in

[10]URL: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html. The URL was last verified on April 26, 2011.

[11]WordNet online: URL: http://wordnetweb.princeton.edu/perl/webwn. The URL was last verified on April 26, 2011.

[12]Merriam-Webster online: URL: http://www.Merriam-Webster.com/. The URL was last verified on April 26, 2011.

(Figure 4.8).

The Mirrors method has partitioned the translations of *gap*N quite plausibly in relation to the *a priori* sense distinctions in WordNet and Merriam-Webster. Out of seven sense partitions (Fig. 4.8), five of them only contain a singleton correspondent. In the first sense partition, the Mirrors method plausibly grouped *avstand* and *tomrom* into one sense through their shared correspondent *void*. In the second partition, *hull* and *åpning* overlapped translationally through *hole* and were therefore grouped together. As the *a priori* sense enumerations in the two lexical

The 1st $t$-image of *gap*N:

{*avstand forskjell hull kløft lakune opphold svelg tomrom åpning*}

Figure 4.7: The first $t$-image of English *gap*N in the ENPC

The sense partitions of the 1st $t$-image of *gap*N:

| | |
|---|---|
| {*avstand tomrom*} | 'void' |
| {*hull åpning*} | 'hole' |
| {*forskjell*} | 'difference' |
| {*kløft*} | 'mountain pass' |
| {*lakune*} | 'lacuna' |
| {*opphold*} | 'time break' |
| {*svelg*} | 'chasm' |

Figure 4.8: The sense partitions of the Norwegian noun *gap*N in the ENPC. Each sense partition is supplied with a suitable English translation in quotes to the right.

resources, the Mirrors singled out the concepts of 'difference' (*forskjell*), 'mountain pass' (*kløft*) and 'time gap' (*opphold*). The *lakune* 'lacuna' sense (which denotes something missing, especially in text) should perhaps have been grouped in the first sense partition, but *lakune* occurred quite rarely and did not have any translational overlap with any other members of the $t$-image of *gap*N. The Mirrors has also captured the rather specific meanings of 'mountain pass' (*kløft*) and 'break' (*opphold*). The 'svelg' sense corresponded with *gap* in the ENPC in the sense of 'the gap between', and hence it might have been grouped together with the 'avstand' partition or with 'forskjell'.

The sparse data issue, preventing some correspondences from being grouped together, is a practical rather than a theoretical problem, and may be expected to improve with the increased availability of parallel data. When using this sense inventory for WSD, it is unfortunate when more senses are generated than what we ideally would prefer; nonetheless, the most crucial issue is that the Mirrors method does not seem to fail at separating the *de facto* contrastive senses.

### 4.4.3 An implementational detail: 'Bag-of-singleton' partitions

A special kind of singleton senses are given 'special treatment' in the current implementation of the Mirrors method by Dyvik, which will henceforth be referred to as *bag-of-singleton partitions*. If a translational correspondent of a word $w$ only occurred with $w$ in the entire corpus, then there is no translational overlap information that can possibly link it to any other words. Normally, this kind of correspondents simply constitute singleton sense partitions of $w$ without any semantic relatedness to any other words. But when introducing automatically word-aligned input to the Mirrors method, more 'noise' was introduced into the sets of translational correspondents; such noise tends to be strongly represented among the mentioned kind of hapax-alignments. Since this may create, in the most extreme cases, a long list of singleton sense partitions that do not have any interesting semantic information about their relation to other words anyway, they are—for practical reasons—grouped together in one sense partition (hence the term 'bag of singletons').

Since the members of a 'bag of singleton' partition do not have any semantic relationship to each other or to other senses, they are unwanted for the purpose of evaluating the translational criterion of the Mirrors method. They are therefore ignored in the WSD experiments of the current thesis. Bag-of-singleton partitions are not formally marked in any way in the Mirrors word bases, and therefore look (*prima facie*) as any other sense. The criterion for detecting them is whether the first member of a sense partition of a lemma $x$ only has the same lemma $x$ as its translational correspondent—if this is the case, then so will be the case with any other members of the same sense partition, since $t$-image members with this property are always grouped together.

While ignoring 'bag of singleton' partitions in WSD exeriments, we do include them when presenting a particular lemma as an example in the text of this thesis. A bag-of-singleton partition will then be represented by a dummy symbol BAG-OF-SINGLETONS. An example of such a bag-of-singleton partition appears in Figure (7.27) (p. 160) in Chapter (7). The first sense partition of the English noun *company*N consisted of the members {*infanterikompani* sementkompani* skiftande* toppleder**}, which are all really singleton sense partitions, and the sense partition is replaced by the dummy symbol BAG-OF-SINGLETONS. They are included in the text simply to visualize the total set of correspondents and how they are treated by the Mirrors method (otherwise a sense may seemingly be missing, causing uncertainty as to why).

### 4.4.4 Semantic fields

Having individuated word senses, the Mirrors method organises the resulting senses into semantic fields. Semantic fields define which senses (lemmas) in a language are semantically related; hence we normally do not find two senses of the same word in the same semantic field. Only senses within the same semantic field can be interrelated by semantic relations such as synonymy and hyponymy.

Keeping English *gap*N as our example, the seven sense partitions from (Figure (4.8) (p. 78)) give rise to seven senses which we denominate as *gap*1, *gap*2, etc. The semantic fields of each sense can be inspected at the Mirrors online[13]; in the present outline we will only show the resulting semantic field for the 'void'-related sense of *gap*1 (Figure 4.9). As the figure shows, the Mirrors has plausibly grouped *gap*1 with *space*2, *void*1 and *emptiness*1.

The Mirrors defines semantic fields on the basis of translation: two word senses $a$ and $b$ in a language $L1$—for instance English *void* and *space*—belong to the same semantic field if their respective first $t$-images overlap. That is to say, they have a correspondent $x$ in language $L2$ in common—for instance, *void* and *space* overlap through Norwegian *tomrom*. Since the relation of sharing a translation is not transitive, there will not necessarily be a shared translation for every pair of members of a semantic field.

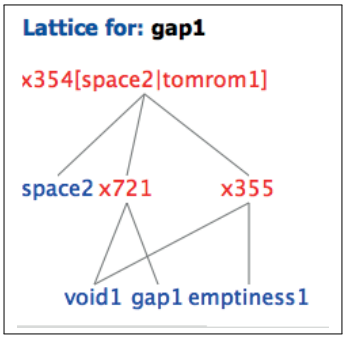



Figure 4.9: Lattice for the first sense of English *gap*N (cf. its sense partitions in Figure (4.8) (p. 78))

Note that the Mirrors caters for ambiguity across languages. The $L2$ word that makes $a$ and $b$ overlap is itself carved into sense partitions; for instance *tomrom* has the following sense partitions: {emptiness gap room space void}, {blank},

[13] URL: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html. The URL was last verified on April 26, 2011.

and {vacuum}. In order to ensure that ambiguity is catered for, the $L1$ words $a$ and $b$ must be in the *same* sense partition of the $L2$ word.

When senses are grouped into semantic fields, this proceeds in parallel between the languages $L1$ and $L2$, such that each semantic field in one language has a counterpart in the other language. For further details, consult (Dyvik, 2005, p. 12 onwards). As we will see in the following, the hierarchical structure of a semantic field is expressed through the assignment of *semantic features*. Semantic fields have a structure because they impose a *subset structure* on each other, since each sense in one field will have its 1st $t$-image as a subset of the other field and vice versa.

### 4.4.5 Semantic features

Having grouped senses into structured semantic fields, the Mirrors method assigns semantic features to all the senses. Specifically, at least one semantic feature is constructed from each sense; additionally senses inherit features from senses higher up in the hierarchy.

In general, semantic features may be described as a theoretical construct which is intended to capture the relatedness between concepts through feature inheritance. For instance, if a concept *a* is a hyperonym of *b* and *c*, then *a* is a more general term. This may be expressed formally through semantic features by letting *b* and *c* share one or more features with *a*, but in addition they have own features that make them more specific than *a* (Figure 4.10). It may be noted that what we call the features is strictly speaking irrelevant; the point is that they provide a formal tool to reveal how concepts are related to each other hierarchially.

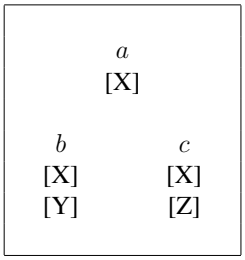



Figure 4.10: Semantic relatedness between concepts, expressed through the sharing of semantic features.

From the translational point of view of the Mirrors method, the basic intuition is that words with wide meanings may be assumed to have many translations (and hence they will be high up in the semantic hierarchy). Words with a narrower meaning, will tend to have fewer translations. Dyvik (2005) uses the

adjective *good* to illustrate how a word with a general meaning will tend to have more translational possibilities than more specific, related word meanings, such as *tasty* (related to one meaning of *good*) or *functional* (related to another meaning of *good*). In the Mirrors method this intuition is modelled by ranking the senses of a semantic field according to how many $t$-images they are members of. Let us begin by considering the senses that we saw in the semantic field of *gap*1 (cf. Figure (4.9)). The hierarchical structure between them is apparent through their semantic features in Fig. 4.11. Each semantic feature is labelled in subscript as either own ($_{own}$) or inherited ($_{inh}$). As can be seen, all senses have the feature $[space2|tomrom1]$ (the own feature of *space*2; inherited by the other senses). In addition to this feature, *void*1 and *emptiness*1 are united through the feature $[void1|tomhet1]$ (an own feature of *void*1), whereas *void*1 has passed on its semantic feature $[void1|avstand3]$ to *gap*1.

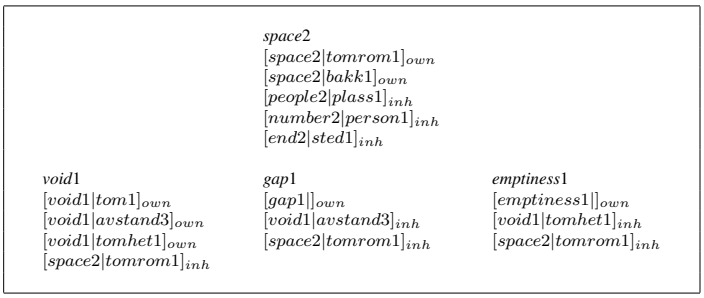



Figure 4.11: The semantic features assigned to the semantic field of *gap*1

The assignment of features is illustrated in detail in Dyvik (2005, fig. 4, p.12). Basically, the method begins by ranking all the word senses in the parallel semantic fields of the languages $L1$ and $L2$, according to their $t$-image memberships (the more $t$-images a sense is a member of, the higher it is ranked). Beginning from the top, a semantic feature is constructed by two elements $[x|y]$ such that the sense $x$ with top rank in $L1$ is paired with the sense $y$ in $L2$ which (i) stands in a translational correspondence with $x$ and which (ii) has the top rank among the $L2$ translational correspondents of $x$. Then the same procedure is run on the top-ranked sense in $L2$, unless this sense happened to be in a translational correspondence with $x$ in $L1$, in which a feature has already been constructed from the given $L2$ sense. In that case we move to the next highest ranked sense in $L2$. Subsequently the procedure continues from the sense with the next highest rank in the $L1$ field, matching it with its $L2$ translational correspondents from the top downwards, always checking whether the translational correspondence is already captured by previous feature assignments, in which case we move further down

the ranking list.

Having constructed a feature, it is then assigned to all word senses that are related: if a word sense $x1$ in $L1$ corresponds to $a1$, $b1$ and $c1$ in $L2$, then all these words receive $x1$'s own feature. In this way, the denotation of a feature $[x1|a1]$ contains the sense $x1$ in $L1$, the sense $a1$ in $L2$, and it also contains the senses that are ranked *lower* in the hierarchy than $x1$ and $a1$. (Whether it is true that senses ranked lower do in fact have a more specialized meaning is, of course, an empirical question.)

### 4.4.6 Deriving thesaurus entries

Lattices of the kind that was illustrated in Figure (4.9) (p. 80) are graphical interpretations of the hierarchy implied by semantic features. Thesaurus entries, on the other hand, allow for some flexibility in how to sift and interpret information from the semantic field. It is in the thesaurus entries that we approximate definitions of synonyms, hyponyms, hyperonyms and related words.


Figure 4.12: Mirrors entry screenshot for the first three senses of the English noun *gap*N, from the Mirrors online database.

A thesaurus entry sums up all the Mirrors senses of a lemma. Figure (4.12)

(p. 83) shows a screenshot of the first three senses of English *gap*N, taken from the Mirrors online database[14]. As we will remember from Figure (4.8) (p. 78), the Mirrors derived seven senses for this lemma, where the latter five were singleton senses without much information.

The thesaurus derivation uses two parameters, namely SynsetLimit and OverlapThreshold. The SynsetLimit is a numerical value that provides a parameterized basis for drawing the line between semantic relations, including synonymy, hyponymy and hyperonymy as well as a more relaxed concept of 'related words'. By varying the SynsetLimit, the user may vary how 'flat' the hierarchy should be (more synonyms or more hyponymy): the higher the value of the SynsetLimit, the more synonyms (a 'flatter' structure), whereas a lower value of the SynsetLimit yields more hyponyms.

The value of the SynsetLimit can either be set manually, or automatically by the formula given in Equation (4.1). Set automatically, the SynsetLimit is computed as a fourth of the number of word senses in the relevant semantic lattice (denoted $N_{senses}$ in the equation below). The minimum value is currently set to $4$ and the maximum value is $20$, because lower or higher values do not seem to yield very interesting outcomes.

$$n = \frac{N_{senses}}{4}, \text{ where } n \text{ is in the range [4,20]} \tag{4.1}$$

The OverlapThreshold parameter (which, as the SynsetLimit, is also a numerical value) defines the division into subsenses. Specifically, the sense partitions remain unaltered, but by varying the OverlapThreshold, the user may experiment with the fine-grainedness within each sense. This is attractive for more complex semantic fields than what we find in any of the senses of *gap*N.

Consider for instance the English noun *call*N in (Figure (4.13) (p. 85)), where we will focus in the first Sense 1, *call*1. The Mirrors defined three main senses, where Sense 1 comprises the notions of 'telephone call, conversation, notice'. The second sense is represented by the Norwegian translation *kallelse*, which was used in the context of a prophetic call[15]. This sense might have been grouped with the first sense, to the extent that a 'prophetic call' or 'request' is akin to a kind of 'notice', but since this is a very specialised meaning of *call* one may also argue that it is plausibly separated as a separate meaning. The third sense encapulates the 'cry, shout' sense. By comparison, three dictionaries (WordNet, Merriam-Webster, Cambridge Dictionary) list a much higher number of senses,

[14]URL: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html. The URL was last verified on April 26, 2011.

[15]The full sentence context has the ENPC sentence ID HB1T: *Til forskjell fra Jesaja, som svarer så ivrig på* **kallelsen***, føler Jeremia (...).* 'Unlike Isaiah, who responds so eagerly to the **call**, Jeremiah (...).'

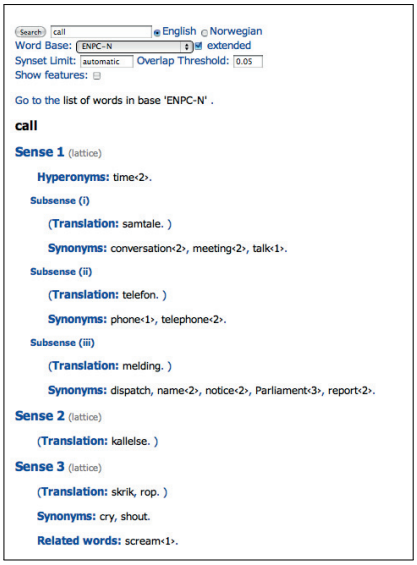



Figure 4.13: Mirrors entry screenshot for the English noun *call*N, from the Mirrors online database. Three main senses; the OverlapThreshold parameter allows to vary the granularity of each main sense.

principally due to fine-grained distinctions in what corresponds to the first sense of the Mirrors.

The semantic field, or lattice, of Sense 1 of *call*N (*call*1) is extremely complex with many members (almost 70 word senses)[16]. Within bigger semantic fields, some senses tend to be more closely related than others, forming, as it were, sub-hierarchies with an internal structure of more and less specific concepts. The OverlapThreshold parameter allows the user to vary the granularity in the carving up of subsenses within a semantic field. As shown in the figure, using the default value for the OverlapThreshold, three subsenses are differentiated, with a shared hyperonym and with their own synonyms. There is one subconcept relating *call*1 to *conversation, meeting, talk*; one concept relating it to *phone, telephone* and one relating it to a 'message' notion including for instance *dispatch, notice, report*. By comparison, if we set the OverlapThreshold to zero (no overlap), then there is only one large sense of *call*1, where all the above mentioned related concepts are

[16]The interested reader may scrutinize the semantic lattice of *call*1 from the Mirrors online word bases: URL: http://maximos.aksis.uib.no:8020/cl/sm/wn-entry.xml. The URL was last verified on April 26, 2011.

lumped together as one list of synonyms.

## 4.5 Conclusion

This chapter has presented the two basic knowledge resources on which the present thesis builds: The Mirrors method, which in turn depends on a preprocessed version of the English-Norwegian Parallel Corpus (ENPC).

The Mirrors method is developed by Dyvik (1998, *inter alia*), and takes translational data at word level as input, and outputs semantic information through the following steps:

- Each lemma is broken down into word senses;
- Each word sense is grouped into semantic fields with other word senses (two senses of the same word normally do not end up in the same semantic field);
- Each semantic field is structured through the assignment of semantic features (with semantic feature inheritance)
- The degree of relatedness between senses may be controlled through the parameters SynsetLimit and OverlapThreshold.

We have seen that one of the strengths of the Mirrors method in the face of sparse data (and, potentially, one of its setbacks as a source of knowledge for statistical WSD) is the fact that the Mirrors method, being a set-theoretic approach, depends on translational overlap and not on statistics: each observation only needs to be recorded once in order to provide useful information. At the minimum, the Mirrors method only needs four translational links (and each link only needs to be recorded once) to conclude that two words $\{x, y\}$ in language $L1$ correspond translationally to a word $a$ in language $L2$, and that $x$ and $y$ must reflect unrelated senses of $a$. Denoting a (symmetric) translational relation as $T$, the following four $T$-pairs will suffice: $T\{a, x\}$, $T\{a, y\}$, $T\{n, x\}$ and $T\{m, y\}$; provided that $n \neq m$ and that $m$ and $n$ only have a translational overlap in $a$. This means that with sparse data, the set-theoretic approach of the Mirrors method may be attractive.

Moreover we have seen that the output of the Mirrors method resembles a thesaurus or a wordnet, as it contains abstract information about word senses and various semantic relations between word senses. The interesting property of the Mirrors method is that whereas hand-crafted lexical resources rest on a variety of considerations when defining lexico-semantic information, the Mirrors method

utilises the *consistent criterion of translations and translational overlap*. It is thus desirable to evaluate how far this criterion takes us.

However, it is not easy to evaluate the Mirrors method, precisely because it investigates a question to which neither science, nor philosophy, has reached a unanimous answer: how can we discover and represent lexical meaning? If there is no real consensus on the delimitation and representation of meaning, how can we plausibly evaluate a resource which pursues a particular approach to discover and represent word meaning? This issue will be discussed in Chapter (5), where we motivate the idea of using WSD as a practical task to evaluate the Mirrors method.

# Part III

# The Mirrors as a knowledge source for WSD

# CHAPTER 5

# METHODS TO EVALUATE A LEXICAL RESOURCE

## 5.1 Chapter introduction

This chapter reviews and discusses methods to evaluate a lexical resource in general (Section (5.2)). We then specify the empirical domain for evaluating the Mirrors method as a lexical resource (Section (5.3)), before we motivate the choice of WSD as a practical evaluation setting in Section (5.4). This section considers previous work specifically aimed at evaluating the Mirrors method within a practical setting in order to show how this has influenced some of the choices made in the current thesis.

## 5.2 Three methods for evaluating word senses

Generally, there are three methods for evaluating word senses which will be treated in order, viz.

- Comparison against a 'gold standard'
- Manual verification
- Validation within a practical NLP task

### 5.2.1 Comparison against a 'gold standard'

Using one or more established lexical resources as a 'gold standard', against which automatically induced senses may be compared, is a well-known approach. A 'gold standard' in the shape of a publicly available lexical resource has the advantage that it is inter-subjectively accessible for re-use and for discussion. In the context of SENSEVAL/SEMEVAL (Section (3.4.2)), for instance, WordNet has become a *de facto* standard sense inventory.

But measuring senses, such as the Mirrors senses, against another sense inventory is simply not very informative, since even high-quality lexical resources diverge in how they carve up senses (cf. the problems with human-crafted lexical resources in Section (3.3.2)). Precision is clearly more important than recall for the Mirrors method, since the main issue concerns, not words that may be missing, but the quality of the information that *is* included. Put bluntly, the absence of a given gold standard word in the Mirrors does not in itself lead us to conclude anything about the quality of what *is* in the Mirrors method.

Thunes (2003) compared the relatedness between senses (synonymy, hyponymy, related words) for 43 English adjectives in the Mirrors word base[1] with their counterpart entries in the Merriam-Webster and the Princeton WordNet (the results are partially repeated in Dyvik, 2009). Thunes's study amply illustrates the problem of quantitative comparisons between sense inventories:

To avoid differences in how semantic relations are defined in the three resources, the comparison did not distinguish between synonyms, hypo- and hyperonyms and related words, but rather took the union of all related words (considering for instance the first sense of *gap*N in (Figure (4.12)), the synonyms *space2, void1* and the related word *emptiness* would have been joined into one set of concepts related to *gap*1). Averaging the full set of 43 adjectives with respect to Merriam-Webster, Thunes (2003) found that the precision and recall of the Mirrors method is 18.5% and 13.5%, respectively. But the evaluation also showed that the intersection of words between Merriam-Webster and WordNet is almost as low as the overlap between the Mirrors method and any of the two gold standard resources.

So if two established lexical resources list semantic relatives of a word which do accord with human intuitions but which hardly intersect, then the contents of established resources may be described as high-quality but incomplete. This, in turn, makes them inadequate as gold standards for precision, since missing words may be related to gaps in the translational input to the Mirrors method and not to the Mirrors assumptions themselves, but the reasons cannot be inferred from counts in themselves.

[1] All translational correspondences for these adjectives were collected manually from the ENPC.

### 5.2.2 Manual evaluation

In order to evaluate abstract senses—as those in the Mirrors method—informants may be consulted under controlled circumstances (cf. Dyvik, 2005, p. 21).

In a more 'concrete' setting, when word senses are somehow represented by corpus instances that are grouped into sets of instances that are taken to express the same sense (e.g. clustering, Section (3.3.3)), humans may manually evaluate whether the membership of each instance in a set is appropriate.

Indeed, the latter kind of manual evaluation could have been an option in the current thesis, since Chapter (7) describes a method to use the Mirrors method as a knowledge source to perform translation-based sense-tagging automatically on the entire parallel corpus ENPC. All corpus instances that were automatically tagged with the same Mirrors sense tags could then be interpreted as a counterpart to data-driven 'clusters' of word instances, which could then be evaluated manually. The use of inter-annotator agreement is then a methodologically sound strategy (cf. Ide et al., 2002), since it diminishes the objection that human judgments are subjective and possibly unstable. It was not chosen to pursue this idea in the current thesis, however, principally because it was desirable to aim for an evaluation framework that allowed to evaluate not only sense distinctions, but also semantic relations between senses. To this end, as we will see, the practical evaluation in an NLP task was deemed more appropriate.

It may be remarked that the Cross-Lingual WSD task in the SEMEVAL-2 (2010)[2] offers a kind of test data that might have provided a very useful independent test set to evaluate the sense distinctions that result from the Mirrors method, as this task, as described in the web page outline, provides manually defined 'clusters' of translations where each cluster points to a distinct meaning of a given target word (this task has also been mentioned and discussed in Chapter (3), pages 34 and 46). This task and the ensuing data have not become available in time for the current project to make use of them, however.

In general, human judgments are clearly problematic since they do not provide a principled point of reference for future controls and discussions. Specifically, they do not provide a useful point of reference for new material, new domains and new languages. Moreover, manual evaluation is costly and time-consuming, which makes it only feasible for smaller-scale evaluation (inter-annotator agreement measures thus become particularly costly, since several annotators are required). Since it is not realistic to repeat evaluation on new texts and domains, the evaluation is typically confined to one selection of text and domain.

But note that the line of reasoning above is not to say that I find qualitative evaluations uninformative. In the setting of evaluation through a practical

[2]URL: http://semeval2.fbk.eu/semeval2.php?location=tasks#T8. The URL was last verified on April 26, 2011.

NLP task (see below), we might find that the use of the Mirrors as a resource does not improve performance significantly, although we may still find reasons for considering the Mirrors assumptions plausible. Among other things, we may find on closer scrutiny that errors in the automatic word alignment of the parallel corpus yield erroneous input to the Mirrors derivation of sense information, which in turn results in erroneous output from the Mirrors. When our focal point of interest is the Mirrors assumptions, it becomes problematic that our evaluation is based on material which is influenced by other factors than what follows directly from the Mirrors assumptions (the output will be influenced by the sheer corpus size and by the quality of the pre-processing of the corpus, such as automated lemmatisation, part of speech-tagging and word alignment). A qualitative evaluation is therefore pertinent in order to isolate the implementational problems of the Mirrors (for instance automatic word alignment) from problems with the Mirrors assumptions *per se*.

### 5.2.3 Practical evaluation in an NLP task

The final approach, and the one taken in the current dissertation, is to evaluate a lexical resource in the context of a practical NLP task. The motivating idea is that a well-defined end-user application may provide a stable framework within which the benefits and drawbacks of a resource or a system can be demonstrated.

#### Related work

There seems to be a growing body of related work where a practical evaluation scheme is advocated.

Resnik (2006, p. 324–325) discusses the idea of using words in a second language as sense labels, and argues that when tagging words with words in another language a connection is made between WSD (sense labelling) and the practical task of machine translation. This is attractive, he argues, since in an end-user system (such as machine learning), the value of a WSD system can more easily be demonstrated.

Agirre and Soroa (2007) discuss evaluation in a practical setting in relation to the word sense induction task at SEMEVAL-1 (2007), suggesting to evaluate the competing word sense induction systems according to their performance in a practical application, such as cross-lingual information retrieval of machine translation. They state that it is "a very attractive idea, but requires expensive system development and it is sometimes difficult to separate the reasons for the good (or bad) performance" (*ibid.*, p. 8).

While it is certainly true that in a practical task it may be difficult to assess the reasons for performance, its benefit is that it provides a stable framework within

which one may study the effect of varying one factor at a time. This principle has for instance been applied for the evaluation of competing WSD systems. Agirre et al. (2007) propose to evaluate WSD systems in the practical task of Cross-Lingual Information Retrieval (CLIR). In a CLIR system a number of choices must be made, so Agirre et al. provide a fixed framework in terms of choices such as translational strategies and which information retrieval system to use; the only thing which is left for the participants to choose is the WSD strategy. A setting is thus created where it is possible to isolate only one factor (namely the effect of each participating WSD system). Agirre et al. state that: "We think that a focused evaluation where both WSD experts and IR experts use a common setting and shared resources might shed light to the intricacies in the interaction between WSD and IR strategies, and provide a fruitful ground for novel combinations and hopefully allow for breakthroughs in this complex area" (*ibid.* p. 2).

Apidianaki (2008) develops a method for sense induction and tests the system in a WSD task (described in Section (4.2.2)). A 'sense' in her system may, similarly to the Mirrors method, be seen as a cluster of translational equivalents. Evaluation is conducted on a parallel corpus where the actual parallel corpus translation of each instance is used as a 'reference translation', and where her system attempts to assign the correct cluster to a corpus instance. The automatic sense-tagging of a corpus instance is deemed to be correct if the system chooses the cluster which contains the reference translation.

Finally, there is a growing body of work that compare the usefulness of different knowledge sources for WSD, which is perhaps the kind of work which is most closely related to the current project (e.g. Ng & Lee, 1996; Stevenson & Wilks, 2001; Yarowsky & Florian, 2002; Specia et al., 2009). The *knowledge sources* in such experiments are commonly taken to be the kind of contextual knowledge used, for instance syntactic relations, selectional restrictions, position-specific information about the local context (collocations), topical information about words that co-occur in a wider context (often termed 'co-occurrences', 'keywords' or a 'bag-of-words'). The basic approach is to isolate each knowledge source by training separate WSD classifiers based on each knowledge source, and then to test each classifier on the same data set. The usefulness of each knowledge source is then measured by its relative contribution compared to the other knowledge sources in the experiment.

For instance, Specia et al. (2009) consider nine knowledge sources, among them collocational knowledge, topical word associations (a bag-of words), syntactic information and selectional restrictions. The contribution of each knowledge source is evaluated on the SemEval-2007 English lexical sample task (65 verbs and 35 nouns, with an average of 222 training examples and 49 test instances). Each knowledge source is evaluated independently by training separate classifiers for each knowledge source, but the data set and the classification al-

gorithm is kept stable. The performance of each classifier is first compared against the most frequent sense baseline; furthermore the effect of classification based on singular knowledge sources is compared against the effect of using all knowledge sources in combination. They also consider the effect of training combinatorial classifiers where all knowledge sources but one are used, thus testing the effect of interactions between classifiers.

In the next, and final section of this chapter, we will outline and discuss the evaluation of the Mirrors method specifically.

## 5.3 The empirical domain for evaluating the Mirrors

There are two aspects of the Mirrors method that lend themselves to evaluation. The first is the sense partitions themselves, which in a practical setting would be taken to be the sense inventory of a word. The second is the semantic relatedness between senses, that is, the wordnet-like aspect of the Mirrors method. These very properties make the thought of WSD spring to mind: Sense-tagging a corpus with Mirrors senses, one could exploit the semantic relations between senses to *enlarge* the training material.

For evaluation in the current thesis, I propose to confine our attention to a tractable lexical sample where the sense partitions accord quite well with intuition. By selecting words with sense distinctions that are as uncontroversial as possible, we provide a framework for focussing, not on how difficult it was (in itself) to learn the senses of the ambiguous target word, but on the effect of varying the contextual information that the classifier learns from (cf. Specia et al., 2009, *inter alia*). The lexical sample and considerations related to it are further discussed in Chapter (8).

Within such a framework, we may then compare the effect of generalising from context words to Mirrors-derived information about context words. This is possible because in the first step, an automatic sense-tagging method is applied to the ENPC, yielding a partially semantically analysed corpus—partially, because the method is only applicable for those instances that have an identifiable translational correspondent (outlined in Chapter (7)). In other words, we will use Mirrors-derived word senses as found through semantic relatedness in the context of a target word. If the quantitative results of such an attempt proves promising, this could be taken to strengthen the Mirrors hypothesis, because it would indicate that the Mirrors generates information that is plausible.

But what will be the empirical content ascribed to terms such as the “quality”, the “plausibility” or the “usefulness” of the Mirrors method? Since the perception

of senses may vary according to what we need them for, we see a need to anchor our judgments to something. The empirical foundation of the Mirrors method, for instance, is in the translational relation, hypothesizing that monolingual semantic information may be derived from it. In that case, the empirical question is not whether the Mirrors method has derived information about a word in language $L1$ which fits its translational properties in $L2$. Rather, the question is whether the translation-based senses and semantic relations in $L1$, given $L2$ as our model, are linguistically motivated *from a monolingual point of view*. The evaluation approach of Apidianaki (2008) (see Section (4.2.2)), which rests on 'reference translations', is then not suitable. Rather, we wish to test if the translation-based sense distinctions of a word are linguistically motivated *from a monolingual point of view*.

The monolingual task of WSD thus offers a suitable evaluation framework for testing the Mirrors as a knowledge source. The proposal is to anchor the notion of sense plausibility to the concept of *learnability* in the context of a machine learning approach to Word Sense Disambiguation (WSD):

1. We expect that there exist discoverable regularities in the usage of a given word (cf. Section (4.2.1)).

2. If the translation-based Mirrors senses prove to be learnable from context, meaning that a classifier is able to retrieve these senses in previously unseen text on the basis of context, then we have shown a correlation between the translational sense criterion of the Mirrors and the context-based sense criterion.

3. We can then argue that two different criteria—the translational criterion of the Mirrors, and the distributional (monolingual) criterion of machine-learning—point to the same sense classification. If two independent criteria support the same sense individuation, it then appears plausible to assume that this individuation reflects a real sense division.

One may immediately object that the two criteria are not in fact independent, since the proposed methodology employs supervised machine learning, which means that the training material is sense-tagged in advance with Mirrors senses. The learning process is then, as it were, biased, since the training instances have been sorted *a priori* by the Mirrors sense classification. But this is precisely our aim, because assigning the Mirrors senses to context examples is the only way in which we can explore the extent to which translation-based senses have contextual correlates.

But what is the justification for saying that our suggested corpus-based evaluation strategy is really more "objectively valid", or informative, than to evaluate

the Mirrors information directly and manually against some dictionary? For instance, if we observe that the machine learning algorithm classifies instances in accordance with the Mirrors senses in, say, 80% of the instances, how are we to interpret this? How can we avoid labelling such results as promising, if that is what we like, or as unpromising, if that is what we consider a beneficial conclusion?

More to the point, following the falsification idea of Popper (1959), we must state the conditions which, at least in principle, would lead us to conclude that the Mirrors method is inadequate as a knowledge source. The "best-known" evaluation methods to assert the "goodness" of a machine learning result in WSD is to measure classification results against a baseline (cf. Chapter (3.4.2)). We may thus state *a priori* that the results of the Mirrors in WSD appear promising if classification performance exceeds chance (if a word is given two senses to choose between in classification, the chance baseline would be 50%) or if it outperforms the approach of simply choosing the most frequent sense (that is, the context-independent *a priori* probability of each sense). This is the evaluation method that is seen in the experiments aimed at comparing the relative contribution of different knowledge sources for WSD (e.g. Specia et al., 2009).

We will also follow Pedersen (2002) in considering whether one knowledge source can resolve instances that alternative knowledge sources could not resolve, and whether it succeeds in resolving the same instances as other knowledge sources. This is particularly well-motivated for us, since we intend to introduce Mirrors-derived information about the *semantic relatives* of the actually co-occurring words. Since a traditional corpus-based knowledge source, such as co-occurrences, is confined to the words that actually co-occur with the target word, it is particularly interesting to see if there is a gain in adding Mirrors-derived information about the semantic relatives: is there actually a gain in adding information beyond what is in the corpus?

## 5.4 WSD to evaluate the Mirrors

### 5.4.1 Previous work: a 'proof of concept' experiment

In a 'proof of concept' experiment (Lyse, 2003, 2006), the practical task of sense-tagging was used to evaluate the Mirrors as a source of lexico-semantic information and to explore the potential of the presented method as an alternative to the manual sense-tagging of corpora for 'supervised' WSD.

This work was done prior to the word-alignment of the parallel corpus, and the evaluation was confined to two nouns for which translational input had been derived manually (the translational correspondences of the noun *rett*N were extracted in Lyse (2003) and those of *tak*N were extracted by Dyvik in the "From

Parallel Corpus to Wordnets"). Since these data stem from manually extracted translational input to the Mirrors method, they illustrate the result when feeding high-quality input into the Mirrors method. The sense distinctions of these two nouns according to the Mirrors method are listed in (Figure (5.1)). The exper-

| | | | |
|---|---|---|---|
| *tak*N | {ceiling roof} | *rett*N | {course} |
| | {grip hold} | | {dish food special supper} |
| | {cover} | | {court justification} |
| | | | {claim entitlement law option order right} |

Figure 5.1: The Mirrors sense-partitions of the nouns *tak*N and *rett*N

iments showed that at least when deriving the translational input to the Mirrors method manually, the kinds of sense distinctions that the Mirrors method outputs appear quite promising. The Norwegian noun *rett*N is intuitively ambiguous between the concept of 'food' and of having a 'claim; right' to something. Although the Mirrors method generated two food-related and two judicially related senses, the actual ambiguities were nonetheless successfully separated. The noun *tak*N is ambiguous between a 'grip, hold' sense and a 'roof, ceiling, cover' sense (the latter is represented by two separate sense partitions, rather than one).

Using the sense partitions as a sense inventory, instances of a target word may be sense-tagged based on its translational correspondent. Sentence (4a) below illustrates an instance of the target word *rett* in the ENPC, in which it corresponds to *dish*. Based on the sense inventory seen in Figure (5.1), the English correspondence *specials* belongs to the 'food'-related sense partition {dish food special supper}; let us for simplicity thus create a sense-tag called DISH. Hence, the TW instance may be assigned the tag DISH, as in (4b).

(4) a. ..smaken av gårsdagens middag med seg til dagens **retter**. (KF1)
..the taste of yesterday's dinner over to next day's specials. (KF1T)

b. ..smaken av gårsdagens middag med seg til dagens ***retter*|DISH**.

With access to a word-aligned corpus, as in the current thesis, it is a trivial procedure to map sense partitions to corpus instances, since the same word alignments provide the basis for deriving sense partitions as well as for assigning senses to corpus instances with the automatic sense-tagger. The precision of the word alignment-based sense-tagger is then intrinsically 100% with respect to the sense partitions.

By contrast, the absence of word-alignment in Lyse (2003, 2006) made it necessary to evaluate the precision of the automatic sense-tagger (since the system did not know which word in the corresponding sentence matched the word to be sense-tagged). Based on the observations from the two target nouns, coverage was moderate, with 58% and 49% for *rett*N and *tak*N, respectively. This

is not surprising, since there are cases where the translator has chosen a non-literal translation of a sentence; without an identifiable correspondent the target word occurrence must be left untagged by the system. As for precision (manually verified) it was found that the main problem for the automated sense-tagging methodology in Lyse (2003, 2006) was practical in nature, namely that the lack of word-alignment meant that the system needed to traverse the entire corresponding sentence: surprisingly often the corresponding sentence would happen to contain a word which did not correspond with the TW but which happened to be a member of a TW sense partition.

A WSD classifier was then trained on the automatically derived training material for the two nouns, using the memory-based TiMBL software (Daelemans et al., 2007). The learnability of the senses was quite promising (*rett*N: 89%, baseline: 73% and *tak*N: 96%, baseline: 87%). The main problems related to the size of retrieved examples and the uneven distribution of senses, which, although unproblematic for the method itself, demands more training data than what is currently available in the ENPC. A problem, in retrospect, with the evaluation setup was also that when associating the Mirrors senses with corpus instances, some of the senses were very low-frequent (less than 10 corpus instances each). Since, in such cases, it is hard to establish if a 'learnability' problem is related to the plausibility of the sense or to its low frequency, they might have been pruned away before classification. In general, this evaluation framework was confined only to the 'learnability' of the target word senses, and did not include the aspect of relatedness between senses which is also a crucial part of the Mirrors method. This has motivated some changes of the evaluation framework in the present thesis.

### 5.4.2 Evaluating word senses and the semantic relatedness between senses

The project idea of the present dissertation is in line with the discussion in Section (3.5.3), where it was argued that added semantic knowledge may be beneficial for WSD. Márquez et al. (2006, p. 206), for instance, conclude that:

> In order to make significant advances in the performance of current supervised WSD systems, we also think that the feature representation must be enriched with a set of features with linguistic knowledge that is not currently available in wide-coverage lexical knowledge bases.

Now, the "classical" WSD approach is to count the correlation rates between unanalysed context words and each sense of the ambiguous word in question, for instance how many times the unanalysed context noun *middag* 'dinner' correlates with *rett*N in the 'food' sense. With the automatic sense-tagger methodology

described in Chapter (7), one may access the senses and the semantic features of context words. Returning to the example based on manual word alignment, we will find that the sense *middag*1 has been assigned two semantic features, given in Figure (5.1). As the figure shows, these two features give rise to several words related to *middag* 'dinner', such as *mat* 'food', *lunsj* 'lunch' and *matvare* 'food'.

We could then hypothesize that any of the word senses that are close in meaning to *middag*1 could have occurred in the same context as *middag*1, even if the other senses are not actually attested in the training corpus for a given ambiguous target word. In other words, we could attempt to generalise from unanalysed, specific context words to classes of words semantically related to the context word. The WSD classifier may then associate a given sense of the ambiguous target word not only with a set of unanalysed word that correlate with the sense, but rather with semantic *classes* of words.

| | | | | | | |
|---|---|---|---|---|---|---|
| Semantic feature: | $[mat1 \mid supper1]$ | | | | | |
| shared by the senses: | mat1 | aftens1 | aftensmat1 | lunsj1 | måltid1 | rett4 |
| | 'food' | 'supper' | 'supper' | 'lunch' | 'meal' | 'dish' |
| Semantic feature: | $[middag1 \mid food5]$ | | | | | |
| shared by the senses: | føde1 | kosthold1 | matvare1 | rett4 | næring2 | |
| | 'nutrition' | 'nutriment' | 'food' | 'dish' | 'nutrition' | |

Table 5.1: The semantic features assigned to middag1 'dinner' by the Mirrors method, and the senses (other than middag1) in the Mirrors wordnet that share this feature.

The abstraction to contextual features could prove to be statistically advantageous since they are shared between words (similarly to the way we extract from conjugated word forms to an abstract lexeme). The approach may thus prove useful as a way of extrapolating from limited training material.

Following the systematic evaluations of knowledge sources discussed in Section (5.2.3), we may test the performance of a WSD classifier with and without added semantic information from the Mirrors method in the training material, while keeping all other factors in the classification setup stable. In view of the previous work in Lyse (2003, 2006), it seems well-motivated to ignore those senses of the target word that are represented less than some heuristic threshold, since it becomes very difficult to assert the level of importance to attach to the statistical findings for very low-frequent items.

If the semantic analysis of context words, provided by the Mirrors, improves the WSD classifier, then we may take this to strengthen the Mirrors hypothesis. In this framework, we may test different ways to determine the relatedness between word senses: should one consider all semantic features related to a given sense of a context word, only those that are general or only those that are passed on to word senses lower in the hierarchy? And should one consider *all* word senses that share a semantic feature (as the illustration in Figure (5.1) implicitly suggests)?

In order to test the plausibility of the *sense distinctions* present in the Mirrors method, we will then run a new series of experiments: If we train on those semantic features that are specific to the word sense present in the context, and if the sense partitions of this context word are plausible, then we should expect much noise to be introduced if we included all semantic features that are associated to the context word in any sense. Seen from the opposite point of view: if we obtain better results when introducing more semantic features than those predicted by the specific sense partition of a disambiguated context word, then this indicates that the sense distinctions are not very plausible.

We thus assume that it is legitimate to make the assumptions stated in Section (5.3): the more closely a word sense inventory reflects the underlying perception of word senses, the better we expect this sense inventory to be 'learnable' in a corpus-driven machine learning framework for WSD. Specifically, we then expect that the ability to 'learn' this sense inventory and classify new target word instances on the basis of it, should at least exceed a baseline of chance or the most frequent sense. We also expect that if two different criteria—the translational criterion of the Mirrors, and the distributional criterion of machine-learning—point to the same sense classification, this may be taken to strengthen the assumption that they reflect an underlying perception of word senses.

## 5.5 Conclusion

We have seen that there are two aspects of the Mirrors method that lend themselves to evaluation. The first is the sense partitions themselves, which in a practical setting would be taken to be the sense inventory of the word in question. This was the only kind of information being evaluated on a smaller scale in Lyse (2003). The second is the semantic relatedness between senses, that is, the wordnet-like aspect of the Mirrors method.

If the Mirrors senses and the semantic relations between them are plausible, then there could clearly be a potential gain in abstracting from context words to their semantic classes: since individual, contextually significant content words are typically low-frequent we may get better statistics by grouping them into semantic classes. Moreover a semantic class allows us to generalise about potential word co-occurrences that are not attested in our training material. On the other hand, there is also a potential loss of information when restricting our attention to only those context words that are sense-tagged automatically: Not all of the words in a common word co-occurrence model are necessarily sense-tagged automatically with the method presented in Chapter (7). We therefore risk to lose some of the actually present corpus information, at the cost of adding information from the Mirrors method.

In order to measure the loss or gain of adding information from the Mirrors method, an experimental framework is suggested where we develop a machine learning classifier for WSD in which training is performed with and without the Mirrors knowledge sources. Such a practical framework where one factor is varied in order to measure the effect of them is well-motivated by previous research.

# CHAPTER 6

# EXPERIMENTAL FRAMEWORK: OUTLINE AND DEFINITIONS

## 6.1 Chapter introduction

This chapter, along with the next two chapters (Chapters (7–8), outline the experimental framework for the practical experiments to be presented in Chapters (9–10). Basically, a series of controlled experiments is proposed, in which the knowledge source to learn from is varied but where we maintain the same experimental framework in terms of the *classification algorithm, data sets, lexical sample and sense inventory*.

This chapter focusses on the component being varied (knowledge sources) in the experimental setup. It also outlines those components that are quite well-known from other WSD experiments. Specifically, Section (6.2) introduces some basic terminology related to the knowledge sources. Section (6.3) outlines and defines the knowledge sources to be tested, namely WORDS (W), SEMANTIC-FEATURES (SF) and RELATED-WORDS (REL-W). The former is a traditional word co-occurrence approach which will represent the 'best-known' classical approach, whereas the two latter represent two kinds of Mirrors-derived information about the word co-occurrences. Section (6.4) presents an overview of how these knowledge sources are applied in experiments. The components of the experimental framework that are kept stable are presented as follows: Section (6.5) motivates and outlines the choice of Naive Bayes as our classification algorithm. Details of the quantitative evaluation in the classification experiments are found in Section (6.6).

Since the lexical sample and the data sets in the current thesis are previously

unknown to the WSD community, some extra space is then devoted to them: The work related to the automated sense-tagging of the ENPC corpus is presented in Chapter (7), whereas Chapter (8) presents the selection of a lexical sample and the development of data sets.

## 6.2 Some basic terminology

In the WSD literature a 'feature' is usually taken to mean a *context feature* representing a typical characteristic of the context of a given word sense. For instance, a co-occurrence feature is a word that often co-occurs with a given sense of the target word.

In order to avoid a terminological confusion between *contextual features* in the WSD experiments and references to *semantic features* from the Mirrors (which may be used as WSD context features), semantic features from the Mirrors will henceforth be referred to as one capitalised word with a hyphen–**SEMANTIC-FEATURE(S)**–or as the shorthand SF.

Correspondingly, we introduce the same notation for the other knowledge types that are used as context features. Three knowledge sources will be compared, viz.:

- WORDS (W)
  Traditional word co-occurrences
- SEMANTIC-FEATURES (SF)
  The Mirrors-derived SEMANTIC-FEATURES found among the traditional word co-occurrences
- RELATED-WORDS (REL-W)
  Based on the denotation of a SEMANTIC-FEATURE, we introduce classes of words that are semantically related

These will be outlined in Section (6.3).

## 6.3 Knowledge sources

This section will define and outline in detail each of the three suggested knowledge sources: WORDS (W), SEMANTIC-FEATURES (SFs) and RELATED-WORDS (REL-Ws).

Each knowledge source will be illustrated through the context available from the example sentence (5), taken from the ENPC. For the convenience of the non-Norwegian reader, an English ambiguous target word *bill*N is used in the example. This target word is to be presented and discussed in Chapter (7), p. 159. In brief, the Mirrors method discovered the following two senses for *bill*N: 'beak' (*bill*2) and 'invoice' (*bill*3). The seemingly missing sense *bill*1 is explained in Chapter (7). The example sentence (5), taken from the ENPC, exemplifies the 'invoice' sense of *bill*N (the three dots at the end of the sentence mark the end of this sentence according to the corpus).

(5) What was it really that they fussed over there in town, in their big flat with all its appliances that regularly broke down (so-called conveniences that demanded both thought and money), meetings, work, appointments, parties, telephones, theatres, ***bills3***, fixed times... (BV1T)

Since the automatic sense-tagger (Chapter (7)) provides a partially semantically analysed corpus—partially because the sense-tagging method only works for lemma instances that have an identifiable translational correspondent—some context words are associated with their appropriate Mirrors sense. The example sentence is automatically sense-tagged as shown in (6), in which the sense-tagged word forms are given in italics with the appropriate number of the Mirrors word sense following it, for instance *fussed*1[1]. The sense-tagged words (surrounding the target word) are: *fussed, town, big, flat, appliances, so-called, conveniences, demanded, thought, work, parties, telephones, theatres*. The not sense-tagged open-class words are *broke, money, meetings, appointments, fixed, times* (*regularly* and *down* are adverbs and are therefore not considered for our purposes).

(6) What was it really that they *fussed*1 over there in *town*2, in their *big*1 *flat*3 with all its *appliances*1 that regularly broke down (*so-called*2 *conveniences*1 that *demanded*1 both *thought*2 and money), meetings, *work*1, appointments, *parties*3, *telephones*2, *theatres*4, ***bills3***, fixed times...(BV1T)

In the following, context words from this example sentence will be used to illustrate the context information that is available with each knowledge source.

[1]The lemmas with Mirrors senses can be inspected at URL: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html. The URL was last verified on April 26, 2011.

### 6.3.1 WORDS (W): A classical word co-occurrence model

The first knowledge source we will apply are traditional word co-occurrences. Formally, the WORD model will be an open-class word co-occurrence model, in which word co-occurrences are defined as being the following:

- document-internal
- belonging to one of the three open word classes that we consider, nouns (N), verbs (V) or adjectives (AJ)
- within a context window of $\pm n$, i.e. we collect the $n$ nearest items on each side of the target word

In general, many kinds of contextual knowledge about words could turn out to be useful for WSD, such as position-specific collocations (which also includes closed-class words) or a selection of the $n$ most statistically informative co-occurrences. There are two important reasons for choosing specifically open-class co-occurrences in a 'bag of words' approach.

First, regarding the choice to use only open-class words: although the ultimate aim is to obtain information that is useful for WSD in general, our present experiments focus very specifically on the effect of abstracting from context words to Mirrors-derived semantic information about these words. Therefore, the WORD (W) model focusses on the kind of context words that is directly relevant for the subsequent consideration of information from the Mirrors method, i.e. open-class (content) words from the word classes nouns, verbs, and adjectives.

Second, the choice to collect co-occurrences in a 'bag of words' approach is motivated by observations from Leacock et al. (1998). The approach of replacing the actually occurring words by the *classes* of semantically related words– the semantic relatives of the actually occurring context words–rests on the same hypothesis that is also pursued in Leacock et al., namely that words closely related to the target word are likely to occur in contexts similar to the target word (cf. Chapter (3)). An observed downside of their so-called *monosemous relatives* approach was that *the semantic relatives of the target word may have other collocations than the target word itself*. For instance, *line* in its formation sense is often followed by a genitival phrase (e.g. *a line of children* whereas the monosemous relative *picket line* misses this collocation. The morale in this for us is that even if a target word sense co-occurs with a context word $x$, and if $x$ is plausibly (linguistically) related to a sense $y$ in the Mirrors method, it is not *necessarily* given that $y$ is a natural co-occurrence of the target word sense. Thus the choice to avoid (position-specific) collocations that include closed-class items should at least tone down this potential problem.

The context window size $n$ is variable; its value will be set for each target word separately following preliminary optimisation experiments. As an example, with a [$\pm 5$] context window the $5$ nearest items on both sides of the target word are collected. Since there are only two open-class items following the target word in the example sentence, $5 + 2$ lemma co-occurrences are collected:

**Example based on example sentence (5):** ([$\pm 5$] context window)
WORDS={*work*N *appointment*N *party*N *telephone*N *theatre*N *fixed*AJ *time*N}

### 6.3.2 Mirrors-derived semantic classes

But how are we to define and implement a class of semantically related words? Recall that a word sense has at least one SEMANTIC-FEATURE that was constructed from this sense (and a sense in the parallel language), i.e. an *own semantic feature* (cf. Chapter (4.4.5), p. 81). Additionally, a sense may inherit SEMANTIC-FEATURES from other senses in the same semantic field (*inherited semantic features*). All word senses that share a SEMANTIC-FEATURE constitute the *denotation* of this SEMANTIC-FEATURE. So since each SEMANTIC-FEATURE unites a set of senses, a semantic class may be modelled as a

- **SEMANTIC-FEATURE (SF) model**:
  Given a word sense that occurs in the context of the target word, register each SF associated to this word sense as an individual entry in the WSD model.

The downside of using SFs is that their potential 'power' for WSD hinges on the static class of word senses in the denotation of a SF: SFs with a large denotation tend to introduce very large classes of 'somehow meaning-related words', in which case the classes may become, literally, quite meaningless. A second alternative is therefore introduced, in which stricter conditions are applied to determine whether two word senses should be regarded as being related. Such sets of word senses will simply be referred to as RELATED-WORDS:

- **RELATED-WORDS (REL-W) model**:
  Given a word sense that occurs in the context of the target word, retrieve its set of RELATED-WORDS, and register each member of this set as an individual entry in the WSD model.

Each of these knowledge types will be treated in order in Section (6.3.3) and Section (6.3.4) below.

### 6.3.3 SEMANTIC-FEATURES (SFs)

The SEMANTIC-FEATURE (SF) model means that a context word that has been sense-tagged automatically is replaced by the SFs associated with this word sense in the Mirrors word bases.

During learning, the frequency of a SF is increased every time any of the senses in its denotation is observed in the context of our target word. Thus, if two word senses $a$ and $b$ share a SF, and if $a$ as well as $b$ co-occur with the target word, their presence is registered jointly through their shared SF. Moreover, if only $a$ (or only $b$) co-occurs with the target word during training, $b$ (or $a$) may still be recognised in a test situation.

During testing, i.e. when attempting to classify previously unseen target word instances, we do not presuppose that the new context words are semantically analysed; that is why we need WSD in the first place. Each context lemma must therefore be looked up in a Mirrors sense lexicon in order to retrieve its set of Mirrors senses, which, in turn, are associated with SFs (recall that two senses of the same word do not normally share semantic features). Each SF is then looked up in the classification model to see if it is known from training.

This model will most likely work best for SFs with a modestly sized denotation, since a large denotation will probably unite too many word senses to be truly informative for a particular sense of the target word. We will illustrate this with two of the sense-tagged context words seen in Example (6) (p. 107), *telephone*2 and *work*1. The word sense *telephone*2 has been grouped into a quite small 'semantic field' (lattice structure) containing only four word senses, viz. *telephone*2 itself and *call*1, *phone*1 and *conversation*2 (Figure (6.1) (p. 110)).

Figure 6.1: Mirrors lattice for the English noun sense *telephone*2

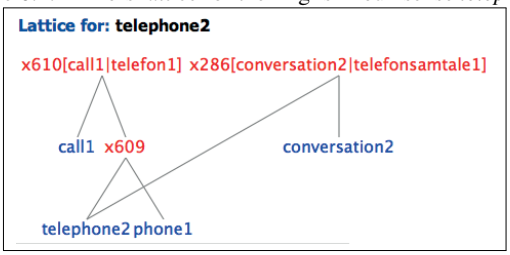



Figure (6.2) lists the SFs (listed to the left) associated with the word sense *telephone*2. The corresponding right column for each SF shows the feature *denotation*, i.e. the word senses that share this SF. The two SFs listed first are the inherited features of *telephone*2 and the third is its own SF. As can be seen, all

three SFs have very small denotations.

Figure 6.2: The SEMANTIC-FEATURES (SFs) associated with *telephone*2, and the denotation of each SF.

**telephone2**

| **SF** | **SF denotation:** |
|---|---|
| $[conversation2\|telefonsamtale1]$ | {*telephone2 conversation2*} |
| $[call1\|telefon1]$ | {*telephone2 phone1 call1*} |
| $[telephone2\|telefonnummer1]$ | {*telephone2 phone1*} |

The *telephone*2 example illustrates how SFs with smaller denotations may give rise to classes of semantically related words that seem plausible. By contrast, SFs that are higher up in a semantic lattice may have quite large denotations. This applies for all three open word classes considered, but the perhaps clearest examples are found with verbs and adjectives. For instance, the verb SFs $[have1|vaere1]$ and $[make1|ta1]$ have 776 and 139 word senses in their denotations, respectively, and the adjective SFs $[great1|stor1]$ and $[large1|liten1]$ have 83 and 60 word senses, respectively.

The SFs of the word sense *work*1 (from Example (6), p. 107) also illustrate large SF denotations. This word sense has two SFs (Figure (6.3) (p. 111)), of which the former is an inherited feature and the latter is its own feature (the full lattice may be inspected online[2]).

Figure 6.3: The SEMANTIC-FEATURES (SFs) associated with the noun sense *work*1, and the denotation of each SF.

**work1**

| **SF** | **SF denotation:** | |
|---|---|---|
| $[business2\|arbeid1]$ | {*assignment1 activity2 age1 business2* ...} | (25 word senses) |
| $[work1\|forhold1]$ | {*affair1 case2 child1 Children1* ...} | (27 word senses) |

Combined, these two SFs give rise to 52 unique word senses (the denotations of these two SFs do not intersect at all), spanning from word senses that are clearly related to *work*1 (*business2, job2*) to more remotely related senses (*industry2, activity2*) and also to word senses where it is hard to see any relatedness at all (*child1, age1*). The latter group may be present in the same semantic lattice as *work*1 due to word alignment errors, although this has not been verified.

This example illustrates that the 'relatedness' between word senses, as defined by the sharing of a SEMANTIC-FEATURE, is very non-flexible since a SF is always

[2]URL: http://maximos.aksis.uib.no:8020/cl/sm/wn-entry.xml. The URL was last verified on April 26, 2011.

connected to a certain denotation: a semantic feature $[x1|y1]$, constructed from $x1$ in language $A$ and $y1$ in language $B$, is passed on to all the translational correspondents of the senses it was constructed from that are ranked lower than $x$ and $y$, respectively. The main function of a system of SF inheritance is to impose a 'general' hierarchical structure on the semantic field, but for our purposes–when focussing on the degree of relatedness based on a specific word sense within a lattice–it may be sound to refine the criteria for considering two word senses as being closely related.

### 6.3.4 RELATED-WORDS (REL-W)

Experimenting with a set of RELATED-WORDS is obviously similar to the way in which concepts such as hyperonyms, synonyms and hyponyms of a sense are determined when deriving thesaurus-like entries in the Mirrors method; the basic aim is to abstract away from some of the detailed information in the full semantic lattices (Chapter (4.4.6), p. 83). As pointed out in Dyvik (2005, p. 14), the sifting of information can be done in more than one way.

The rest of this section is divided into two parts. The first part takes as its starting point how 'relatedness' between senses is determined in the Mirrors method (Dyvik, 2005, p. 16-17), in order to see why and how they may be adjusted. This forms the motivating basis for defining, in the second part, a set of criteria to determine RELATED-WORDS.

#### 'Semantic relatedness' in the Mirrors method

The Mirrors method generates several different semantic relations for a word sense. By contrast, our main concern is rather whether there *is* a relation of (close) similarity between two word senses, and not so much whether such a relation pertains to for instance synonymy or hyponymy. Bear in mind that a class of RELATED-WORDS is to be determined with respect to a particular word sense that actually occurs in the context of a target word. Based on the observed, contextual presence of a word sense $a$, we aim to approximate a selection of RELATED-WORDS that could, conceivably, occur in the same (or similar) contextual environment as $a$ itself.

In other words, we want to tentatively neutralise the difference between the semantic relations that are currently being explored in the Mirrors method. Moreover, it is also necessary to sharpen the conditions that determine a class of RELATED-WORDS, since neutralising the difference between the Mirrors-derived semantic relations alone would result in a potentially large union of synonyms, hyponyms and other related words.

The conditions being currently explored in the Mirrors method are given in the three definitions below. Let $a$ be a word sense that was found in the context of an ambiguous word, and let $b$ be a word sense in the same semantic lattice as $a$. The Mirrors definitions are then as follows.

**Hyperonyms and hyponyms:** Hyperonymy and hyponymy are converse relations, so if $a$ is a hyponym to $b$, $b$ is a hyperonym til $a$. $b$ is a hyperonym to $a$ (and $a$ is a hyponym to $b$) if $a$ inherited $b$'s own SF and if the denotation of this SF is *higher than* the SynsetLimit.

**Synonyms:** $a$ and $b$ are synoymous if
(i) $a$ inherited $b$'s own SF, or vice versa, and if the denotation of this SF is *lower than or equal to* SynsetLimit, or if
(ii) $a$ and $b$ share at least two SFs where the denotation of each SF is *lower than or equal to* SynsetLimit.

**Close-related words:** $a$ and $b$ share exactly one SF where the denotation of each SF is *lower than or equal to* SynsetLimit.

If we break this down into single propositions, the interplay between the sharing of SEMANTIC-FEATURES (SFs) and the denotation size (abbreviated *denot*(SF)) is shown schematically in Table (6.1).

Table 6.1: The Mirrors definitions of semantic relations: a schematic overview given two word senses $a$ and $b$

| | | | | |
|---|---|---|---|---|
| $a$ inherited $b$'s own SF | AND | denot(SF) $>$ SynsetLimit | $=>$ | hyperonym($b$,$a$) |
| $b$ inherited $a$'s own SF | AND | denot(SF) $>$ SynsetLimit | $=>$ | hyponym($b$,$a$) |
| $a$ inherited $b$'s own SF | AND | denot(SF) $\leq$ SynsetLimit | $=>$ | synonym($b$,$a$) |
| $b$ inherited $a$'s own SF | AND | denot(SF) $\leq$ SynsetLimit | $=>$ | synonym($b$,$a$) |
| $a$ and $b$ share at least two SFs | AND | denot(SF) $\leq$ SynsetLimit | $=>$ | synonym($b$,$a$) |
| $a$ and $b$ share one SF | AND | denot(SF) $\leq$ SynsetLimit | $=>$ | 'close-related'($b$,$a$) |

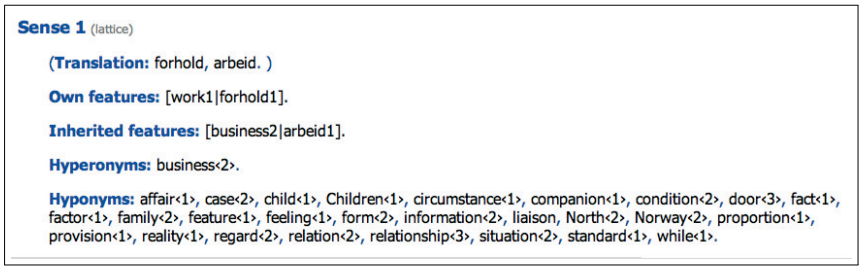
As Table (6.1) shows, the difference between synonyms and hyponyms/hyperonyms is adjusted through the parameter *SynsetLimit*: the higher the value of the SynsetLimit, the more synonyms. The computation of the automatically set SynsetLimit is shown in Chapter (4.4.6), p. 83. (The second parameter when generating a wordnet-like entry, *OverlapThreshold*, determines subsenses and is not relevant here).

Following these definitions, Mirrors entries for the two example context senses *telephone*2 and *work*1 are generated as seen in Figure (6.4) (p. 114) and Figure (6.5) (p. 114)[3]. The SynsetLimits of the two words are set automatically to 4 and 20, respectively. Recall from Figure (6.1) (p. 110) that *telephone*2 is a member of a very small lattice of only four word senses, which means that no denotations exceed the SynsetLimit. Therefore this word sense only has synonyms

[3]These Mirrors entries were generated with the following settings: Word base=ENPC-N, OverlapThreshold=0 (i.e. no sub-senses are generated), SynsetLimit=automatic.

Figure 6.4: Mirrors-defined entry for the English noun sense *telephone*2, SynsetLimit=automatic, OverlapThreshold=0


Figure 6.5: Mirrors-defined entry for the English noun sense *work*1, SynsetLimit=automatic, OverlapThreshold=0




and no hypo-/hyperonyms in its Mirrors entry (cf. the conditions in Table (6.1)). By contrast, *work*1 has no synonyms, because the relevant SFs had a denotation $>$ SynsetLimit. It has one hyperonym (*business*2, which passed on its own SF to *work*1) and several hyponyms (word senses that inherited *work*1's own SF, which had a denotation higher than the SynsetLimit).

### A definition of RELATED-WORDS

Based on the original conditions being currently explored in the Mirrors method, we may now define some tentative conditions to determine classes of RELATED-WORDS. Since the Mirrors method is not well-known compared to the established, hand-made wordnets, it may be instructive to show the transition from the Mirrors definitions to the final conditions for determining RELATED-WORDS in two steps. First, a 'draft' definition is given which is a 'near at hand' suggestion given the discussion of the original Mirrors definition above. The then resulting classes of RELATED-WORDS for a selection of word senses will then serve to explain the final

definition of a class of RELATED-WORDS (which is slightly more complex than the draft).

**RELATED-WORDS** (draft definition):

Given a context sense $a$, which gives rise to a set of SEMANTIC-FEATURES (SFs), a denotation sense $b$ is regarded as a RELATED-WORD to $a$ if at least one of the two following conditions applies:

i. $b$ has inherited one of $a$'s *own* SFs or $a$ has inherited one of $b$'s *own* SFs (regardless of the size of its denotation);

ii. $a$ and $b$ share at least two SFs that are below or equal to the (automatically set) SynsetLimit.

The difference between the Mirrors definitions and the draft definition is schematically shown in Table (6.2), which is the counterpart to the schematic overview of the Mirrors definitions in Table (6.1) (p. 113). As can be seen, the draft definition is more restrictive in that 'close-related' words are omitted. Apart from that, the same words are included in the draft definition as with the Mirrors definitions, except that there is no longer a tentative distinction between hyperonyms, hyponyms and synonyms. Condition (i) in the draft definition neutralises, in effect, the original Mirrors distinction between hypero-/hyponymy and synonymy since it disregards the denotation size of the own features that are passed on to either $a$ or $b$. Condition (ii) maintains the second part of the original synonymy definition: if $a$ and $b$ share at least two of $a$'s SFs, and if the denotations of these SFs are relatively small (below the SynsetLimit), then it is very likely that $a$ and $b$ are closely related.

Table 6.2: The draft definition of RELATED-WORDS: a schematic overview given a contextual word sense $a$ and a candidate RELATED-WORD $b$

| | | | | |
|---|---|---|---|---|
| $a$ inherited $b$'s own SF | | | => | RELATED-WORDS($b$,$a$) |
| $b$ inherited $a$'s own SF | | | => | RELATED-WORDS($b$,$a$) |
| $a$ and $b$ share at least two SFs | AND | denot(SF) $\leq$ SynsetLimit | => | RELATED-WORDS($b$,$a$) |

Following this draft definition of RELATED-WORDS, Figure (6.6) (p. 117) shows some examples of word senses and their resulting RELATED-WORDS. When discussing these word senses, the contextual word sense may be referred to as $a$ whereas $b$ denotes any word sense that is found to be related to $a$.

Basically, the examples indicate that the draft definition of 'relatedness' is not

strict enough. This conclusion is based on a larger set of word senses than those included in Figure (6.6); however, for reasons of space it was decided to include only a few illustrative examples. The examples include the two example word senses already seen, viz. *telephone*2 and *work*1, but also include some more word senses—verbs (V), adjectives (AJ) and nouns (N)—from different sentence contexts of *bill*N. Discussing each sense in detail would take this chapter a bit too far; therefore only two word senses will be analysed in some detail in order to show the lines of reasoning, viz. the verb sense *give*1 and the adjective sense *long*1. These illustrate well where there is a gain in restricting the conditions for including a word sense in the RELATED-WORD. The remaining senses in Figure (6.6) will only be referred to in terms of what they exemplify.

The most 'extreme' example of the need for stricter 'relatedness' conditions is the verb sense *give*1, which has $80$ word senses in its class of RELATED-WORDS. It has three inherited SFs, i.e. three word senses are included in the class of RELATED-WORDs because they passed on their own SFs to *give*1, viz. *be*1, *get*1 and *have*1. The own SFs of these three word senses have denotations $> 20$, and it is hard to see a clear meaning similarity between them and *give*1 since their meaning potential is so wide. The rest of the word senses in the RELATED-WORD class are included because *give*1 passed on its own SF to them. The own SF of *give*1 has, in other words, a quite large denotation. When reviewing the word senses that inherited *give*1's own SF, it is hard to find a convincing meaning relation between then.

So from this word sense example one might hypothesise that it is beneficial to exclude $b$-word senses that inherited the own SF of the $a$-sense if its denotation is quite large. Also, since none of the three senses that passed on their own SF to *give*1 were clearly related to $a$, one might also consider excluding $b$ senses that passed on to sense $a$ an own SF with a denotation $>$ SynsetLimit.

We will also consider the adjective sense *long*1 in some detail. This word sense co-occurred with *bill*N in the sense of physical length (*a long bill*|BEAK). The three inherited SFs originated (on the English side) from *strange*1, *large*1 and *great*1; the two latter are quite plausibly related to a physical sense of *long* whereas the relationship between *long*1 and *strange*1 is perhaps less convincing. All three SFs had quite large denotations, comprising $37$, $61$ and $84$ word senses, respectively. So with regard to $b$-senses that passed on their own feature to the contextual word sense $a$, this example counters the observation with *give*1 above; with this adjective it appears *beneficial* to include word senses that passed on its own SF to *long*1, even when their denotation size is quite large. With regard to $b$-senses that inherited $a$'s own SF, on the other hand, this adjective sense shows the same tendency as with *give*1 above: the vast majority of word senses in the class of RELATED-WORDS is word senses that inherited the *own* SF of *long*1, and most of them are–intuitively–not particularly close in meaning to *long*1. So it seems that it is not necessarily fortunate to include the full denotation of a SF that was

Figure 6.6: Some word senses in the context of *bill*N and their RELATED-WORDS (REL-Ws) following the 'draft definition'

| **Word sense** | **REL-W** | |
|---|---|---|
| *give*1 (V) | {*address1 administer1 be1 bottle1 celebrate1 clear1 close1 clutch1 come1 conceal1 conduct1 confine1 consider1 continue1 control1 cope1 could1 cover1 cultivate1 decide1 defend1 deliver1 die1 divorce1 do1 endure1 find1 fold1 follow1 forced1 forget1 get1 give1 guard1 gulp1 hand1 have1 hide1 hold1 keep1 last1 lecture1 let1 lie1 like1 limited1 linger1 live1 maintain1 obscure1 observe1 occur1 outline1 outlive1 permit1 prevail1 protect1 provide1 rake1 release1 reside1 run1 settle1 shall1 share1 should1 show1 sleep1 smile1 stand1 stop1 struggle1 suppose1 sustain1 threaten1 throw1 undertake1 wake1 work1 would1*} | (80 word senses) |
| *explain*1 (V) | {*achieve1 argue1 exist1 explain1 handle1 have1 imagine1 inform1 please1 prosper1 resist1 say1 succeed1 whistle1 wipe1*} | (15 word senses) |
| *catch*1 (V) | {*be1 bring1 carve1 catch1 clutch1 constitute1 enable1 engage1 experience1 gear1 get1 grab1 grasp1 grip1 happen1 hear1 hit1 make1 penetrate1 protect1 reach1 score1 seeing1 seize1 snatch1 stretch1*} | (26 word senses) |
| *crack*1 (V) | {*burst1 crack1 feel1 snap1*} | (4 word senses) |
| *long*1 (AJ) | {*able1 absolute1 clear2 cold1 complete1 dead2 early1 entire1 equal1 female1 fine1 full2 general1 great1 large1 last1 long1 much1 new1 open2 other1 reliable1 same1 solid1 strange1 thorough1 total1 very1 white2 whole1 wide-open1 wide1 wild2*} | (33 word senses) |
| *various*1 (AJ) | {*different1 multiple1 strange1 various1*} | (4 word senses) |
| *work*1 (N) | {*affair1 business2 case2 child1 Children1 circumstance1 companion1 condition2 door3 fact1 factor1 family2 feature1 feeling1 form2 information2 liaison1 North2 Norway2 proportion1 provision1 reality1 regard2 relation2 relationship3 situation2 standard1 while1 work1*} | (29 word senses) |
| *food*1 (N) | {*age1 bit2 date1 Day1 day2 daybreak1 feeding2 food1 King1 light4 Monday2 nourishment1 pair2 past1 Sami2 Sea2 stage2 today1 year2*} | (19 word senses) |
| *telephone*2 (N) | {*call1 conversation2 phone1 telephone2*} | (4 word senses) |
| *animal*2 (N) | {*animal2 being3 creature1 organism2*} | (4 word senses) |
| *hole*2 (N) | {*gap2 hole2 pit1*} | (3 word senses) |

passed on to other senses.

We will not proceed with commenting each of the remaining word senses in detail, but only state briefly: of the considered example word senses (verbs, adjectives, nouns), only two word senses–*telephone*2 and *animal*2– met condition (ii) (cf. the draft definitions on p. 115), i.e. the condition of sharing at least two SFs with a denotation $\leq$ the SynsetLimit. Based on this condition, *telephone*2 was plausibly associated with the word sense *phone*1 and *animal*2 with *organism*1. There is thus no obvious reason to omit this condition.

Regarding the condition with word senses $b$ that pass on their own SFs to $a$ ($b$ being 'hyperonyms' to $a$ in the Mirrors definitions) the following observations can be made: when $b$'s own SF has a relatively low denotation, $b$ is generally plausibly related to $a$.

**Examples:** The word sense *crack*1 was associated to *burst*1 and *snap*1; *various*1 was associated to *different*1; *telephone*2 was associated to *conversation*1 and *call*1; *animal*2 was associated to *being*1 and *creature*1.

There are of course exceptions, which in Figure (6.6) (p. 117) is illustrated with the sense *food*1. This sense inherited three SFs that originated (on the English side) from *bit*2, *age*1 and *day*2, which had denotation sizes of, respectively, 7, 15 and 15. Of these, *bit*2 might be accepted as a semantic relative to *food*1, whereas the two latter (which also had the highest denotations) do not seem to bear any obvious relationship to *food*1.

As for word senses $b$ that are included because they inherited $a$'s own SF ('hyponyms' in the Mirrors definitions), it seems to be a general tendency that when the denotation of $a$'s own SF is low, the $b$s are quite plausible.

**Examples:** *various*1 was associated with *multiple*1; *food*1 was associated with *nourishment*1 and *feeding*2; *telephone*2 was associated with *phone*1[4], *gap*2 was associated with *hole*2 and *pit*1,

By contrast, when the denotation of $a$'s own SF is relatively high, the resulting $b$s are usually not very strongly related.

**Examples:** *long*1 passed on its own SF to 30 word senses; *work*1 passed on its own SF to 28 word senses, *food*1 passed on its own SF to 15 word senses of which only two word senses were clearly related.

Based on these observations, the conditions in the draft definition are refined as follows:

[4]The two senses *telephone*2 and *phone*1 additionally shared to SFs that were below the SynsetLimit, so they met conditions (i) as well as (ii) of the draft definition.

**RELATED-WORDS** (final definition):

Given a context sense $a$, which gives rise to a set of SEMANTIC-FEATURES (SFs), a denotation sense $b$ is regarded as a RELATED-WORD to $a$ if at least one of the three following conditions applies:

i. $a$ inherited $b$'s own SF, and this SF has a denotation $\leq$ SynsetLimit.

ii. $b$ inherited one of $a$'s own SFs, and this SF has a denotation $\leq$ SynsetLimit.

iii. $a$ and $b$ share at least two SFs that are below or equal to the SynsetLimit.

This definition may also be presented schematically as in Table (6.3) (p. 119). In comparison to the tentative Mirrors definitions (Table (6.1) (p. 113)), the final RELATED-WORD definition is more restrictive in that we omit 'close-related' words as well as words that fall into the category of 'hyponyms' as well as 'hyperonyms'.

The SynsetLimit determines whether we include a $b$ sense that passed on its own SF to $a$ or that inherited $a$'s own SF: a low SynsetLimit value means that very few word senses are included, and vice versa. Indeed, Dyvik (2005) points out that the current SynsetLimit definition is still at an experimental stage in the sense that the effect of varying its value has not been systematically studied. It may therefore be of some interest to incorporate different SynsetLimit values into our experiments. A high SynsetLimit value will lead to the inclusion of word senses that would be pruned away (as 'hyponyms' and 'hyperonyms', according to the tentative Mirrors definitions) with a lower SynsetLimit value.

Table 6.3: The final definition of RELATED-WORDS: a schematic overview given a contextual word sense $a$ and a candidate RELATED-WORD $b$

| | | | | |
|---|---|---|---|---|
| $a$ inherited $b$'s own SF | AND | denot(SF) $\leq$ SynsetLimit | => | RELATED-WORDS($b$,$a$) |
| $b$ inherited $a$'s own SF | AND | denot(SF) $\leq$ SynsetLimit | => | RELATED-WORDS($b$,$a$) |
| $a$ and $b$ share at least two SFs | AND | denot(SF) $\leq$ SynsetLimit | => | RELATED-WORDS($b$,$a$) |

From the these tentative, final definitions to determine a class of RELATED-WORDS, the word senses discussed in Figure (6.6) (p. 117) now get new classes of RELATED-WORDS, as shown in Figure (6.7). With the final definitions, the RELATED-WORDS that were already quite plausible in Figure (6.6) mostly remain (*crack1, various1, telephone2, animal2, hole2*), but the conditions are clearly restricted to yield more precise classes with regard to those word senses that had too many word senses in Figure (6.6) (*give1, explain1, catch1, food1, various1*). Some of the intuitively 'good' related word senses are inevitably lost, for instance

Figure 6.7: Some word senses in the context of *bill*N and their RELATED-WORDS (REL-Ws) following the 'final definition' (SynsetLimit=automatic)

| **Word sense** | **REL-W** | |
|---|---|---|
| *give*1 (V) | {*give1*} | (1 word sense) |
| *explain*1 (V) | {*argue1 explain1 handle1 imagine1 inform1*} | (5 word sense) |
| *catch*1 (V) | {*carve1 catch1 clutch1 engage1 gear1 grab1 grasp1 grip1 happen1 hear1 protect1 reach1 seize1 snatch1 stretch1*} | (15 word sense) |
| *crack*1 (AJ) | {*burst1 crack1 snap1*} | (3 word senses) |
| *long*1 (AJ) | {*long1*} | (1 word senses) |
| *various*1 (AJ) | {*different1 multiple1 various1*} | (3 word senses) |
| *work*1 (N) | {*work1*} | (1 word senses) |
| *food*1 (N) | {*age1 bit2 day2 feeding2 food1 nourishment1*} | (6 word senses) |
| *telephone*2 (N) | {*call1 conversation2 phone1 telephone2*} | (4 word senses) |
| *animal*2 (N) | {*animal2 being3 creature1 organism2*} | (4 word senses) |
| *hole*2 (N) | {*gap2 hole2 pit1*} | (3 word senses) |

*long*1 lost *great*1 and *long*1, because these two word senses passed on their own SF to *long*1 and these own SF had a denotation above the SynsetLimit. For the same reason, *work*1 lost *business*2.

## 6.4 Experimental setup: overview

Having introduced the definitions of our three main kinds of knowledge sources, this section will outline how these knowledge sources will be tested in order to shed light on the viability of the Mirrors as a knowledge resource.

### 6.4.1 Comparing and combining knowledge sources

In Chapter (9) we will compare the different knowledge sources (abbreviated KS) using otherwise identical experimental settings in the following set of experiments (abbreviated EXP) per target word:

**EXP1** KS=WS: The [$\pm$ $n$] nearest WORDS.

**EXP2** KS=SFS: The SEMANTIC-FEATURES (SFS) derived from those words in EXP1 that were automatically sense-tagged.

**EXP3** KS=REL-WS: The RELATED-WORDS (REL-W) derived from those words in EXP1 that were automatically sense-tagged.

**EXP4** Combined KS=W + SF + REL-W

EXP1 is the natural starting point of our experiments, representing a traditional

WORD co-occurrence model for WSD. This model is taken to provide a kind of experimental baseline, word co-occurrences being the 'best-known' method among the proposed experiments of the current thesis. The natural next step is to perform a series of controlled experiments aimed at measuring the loss or gain in replacing context words by their Mirrors-derived SEMANTIC-FEATURES or RELATED-WORDS (EXP2 and EXP3). The last experiment of Chapter (9) combines the Mirrors information with the actually attested corpus WORDS; EXP4 combining all three KSs.

Note that since not all context words are sense-tagged, not all context words *have* Mirrors-derived, sense-specific information. Figure (6.8) shows three different co-occurrence models for the English target word *bill*N, based on Example (5) (p. 107) with a co-occurrence window of $[\pm 5]$: the $[\pm 5]$ nearest WORDS, the SEMANTIC-FEATURES of these words in the relevant word sense (where available) and the RELATED-WORDS of these words in the relevant word sense (where available). As can be seen, three of the $[\pm 5]$ nearest lemmas do not contribute in the SF and REL-W model because they were not sense-tagged (*appointment*N, *fixed*AJ and *time*N).

These experiments address the following research questions:

- EXP1: how well may a traditional WORD classifier be expected to perform, given our specific data sample, sense inventory and classification algorithm?
- Replacing context words with Mirrors-derived information (EXP2, EXP3): do they display complementary benefits with respect to EXP1?
- Adding Mirrors-derived information (EXP4): What is the loss or gain in adding paradigmatic information from the Mirrors method? Does added information in fact lead to more confident and more correct classifications? (or does Mirrors-derived information introduce more noise?)

The exact loss of information cannot be analysed in much detail, since our main emphasis lies on the properties of the Mirrors method for WSD–the information loss that is caused by missing word-alignment, by contrast, is rather an accidental property of the parallel corpus and the automatic word-aligner.

### 6.4.2 Measuring the loss or gain in adding information from the Mirrors method

Chapter (10) presents a more purely theoretical evaluation which measures directly the loss or gain in abstracting from context words to Mirrors-derived information by isolating the context words that are sense-tagged. That is, the W model

Figure 6.8: Three different co-occurrence models for the English target word *bill*N, as outlined for EXP1, EXP2 and EXP3. Based on Example (5) (p. 107), with a co-occurrence window of $[\pm 5]$

| WORD (W) | SEMANTIC-FEATURE (SF) | RELATED-WORDS (REL-W) |
|---|---|---|
| *work*N | $[business2\|arbeid1]$<br>$[work1\|forhold1]$ | {*business2 work1*} |
| *appointment*N | – | – |
| *party*N | $[year2\|parti1]$<br>$[side2\|side1]$<br>$[party3\|selskap1]$<br>$[party3\|gruppe1]$ | {*party3 side2 year2*} |
| *telephone*N | $[conversation2\|telefonsamtale1]$<br>$[call1\|telefon1]$<br>$[telephone2\|telefonnummer1]$ | {*call1 conversation2 phone1 telephone2*} |
| *theatre*N | $[theatre4\|teater1]$ | {*theatre4*} |
| *fixed*AJ | – | – |
| *time*N | – | – |

(and the corresponding SF and REL-W model) is based on the $n$ nearest words that were sense-tagged. The set of exeriments may be presented as follows:

**EXP5** WS: The $[\pm n]$ nearest WORDS that were sense-tagged.

**EXP6** SFS: The SEMANTIC-FEATURES (SFS) derived from all the actually occurring context lemmas in EXP5.

**EXP7** REL-WS: The word sense associated to each context word in EXP5 together with the RELATED-WORDS of each such context word sense.

**EXP8** UNION REL-WS: The union of *possible* Mirrors word senses (irrespective of which is predicted according to the automatic sense-tagging) for each context word in EXP5 together with the RELATED-WORDS associated to each such word sense.

Since we expect that SEMANTIC-FEATURES may lead to a too general kind of semantic classes, one must perhaps expect that the results could come out worse than with WORDS. This is because the generalisations may be too broad, for instance a general verb SF such as $[have1|vaere1]$ is not likely to occur only with one sense of the target word. Whether or not SFs are in fact too often too broad remains to be seen in the experiments, but it would not be surprising.

As regards RELATED-WORDS, on the other hand, the definitions have been designed to avoid very general related word senses. Comprising the union of the actually occurring context words and the RELATED-WORDS of these context words, the theoretical 'worst case' should be that there is no difference between a WORD-

based and a RELATED-WORD-based classification; in the best case there could be a gain because the classifier learns about more words than what is actually attested in the training contexts of the target word. If the Mirrors-derived semantic information results in a performance loss, however, it will be very interesting to consider more closely those instances that came out less well in the RELATED-WORD-based model. Specifically, if the use of RELATED-WORDS give poorer result than the context lemmas themselves, there is reason to question the plausibility of the Mirrors-derived information.

Returning again to sense-tagged version of the example sentence, in Example (6), a context window of [$\pm 5$] based on only *sense-tagged* context words yields the information shown in Figure (6.9) (p. 123). The lemmas in the WORD model are now collected from a wider window than what we saw with the WORD model in EXP1. Since we only collect the $n$ nearest from each side, 5 lemmas are collected from the left side of the target word whereas no lemmas to the right of the target word are sense-tagged. (In the real experiments, however, we do not confine our attention only to sentence-internal information).

Figure 6.9: Three different co-occurrence models for the English target word *bill*N, as outlined for EXP5, EXP6 and EXP7. Based on *sense-tagged* lemmas in Example (6) (p. 107), with a co-occurrence window of [$\pm 5$]

| WORD (W) | SEMANTIC-FEATURE (SF) | RELATED-WORDS (REL-W) |
|---|---|---|
| *thought*N | [$consideration1$\|$omtanke1$]<br>[$idea1$\|$tanke1$]<br>[$thought2$\|] | {*consideration1 idea1 thought2*} |
| *work*N | [$business2$\|$arbeid1$]<br>[$work1$\|$forhold1$] | {*business2 work1*} |
| *party*N | [$year2$\|$parti1$]<br>[$side2$\|$side1$]<br>[$party3$\|$selskap1$]<br>[$party3$\|$gruppe1$] | {*party3 side2 year2*} |
| *telephone*N | [$conversation2$\|$telefonsamtale1$]<br>[$call1$\|$telefon1$]<br>[$telephone2$\|$telefonnummer1$] | {*call1 conversation2 phone1 telephone2*} |
| *theatre*N | [$theatre4$\|$teater1$] | {*theatre4*} |

### Testing the quality of the Mirrors sense distinctions

As a final part of the theoretical evaluation in Chapter (10), the quality of the Mirrors sense distinctions will be tested in EXP8, which uses the UNION of RELATED-WORDS from all possible Mirrors senses of a context word.

In this experiment the sense partitions are ignored, i.e. Mirrors-derived semantic information from all the possible senses of a context word are considered.

The rationale is that since the Mirrors method tends to generate more sense distinctions than we intuitively find desirable, it may happen that more than one of the (Mirrors-) senses of a context word could valuably contribute to characterise the particular context of a target word. Specifically, if we ignore the Mirrors sense distinctions (reflected in sense-specific SEMANTIC-FEATURES or RELATED-WORDS) and instead take the union of Mirrors-derived information about a context word in *any* of its senses, we would expect that if the Mirrors senses are plausible, WSD classification based on sense-specific information should outperform (or equal) classification performance when training on the union of information of all the possible senses of a context word. From the opposite angle, if the learnability of sense-specific information of a context word is outperformed by replacing them with the union of information, this would indicate that the information based on the union is more informative that the sense-specific information. It must be expected that when considering the union of Mirrors-derived information, contextually irrelevant words are also included; one may however hope that irrelevant words are cancelled out statistically.

If the *union* of semantic information about a context word is more useful than the sense-specific information, this would indicate that too much relevant information is spread across several senses, i.e. that the Mirrors senses are not adequate for WSD purposes.

This experiment could in principle be carried out by considering, not only the *sense-tagged* context words, but any context word that has an entry in the Mirrors word bases; its semantic features in every possible sense could then be retrieved. But it may be interesting to consider this experiment in relation to the corresponding sense-specific information in Chapter (10); therefore the union of Mirrors-derived information will also be based on the $n$ nearest words that were sense-tagged (although we then disregard the particular sense predicted by the automatic sense-tagger in that particular context).

So the main questions are thus:

- Chapter (10): Given only sense-tagged context words, what is then the loss or gain in information when we replace the lemmas in a WORD model with Mirrors-derived information about each lemma?

- Does classification actually improve when abstracting from contextual lemmas to classes of semantically related words (expressed as SEMANTIC-FEATURES or as RELATED-WORDS)?

- What is the quality of the Mirrors sense distinctions?

## 6.5 Naive Bayes classifier

### 6.5.1 Motivating the choice of algorithm

There is a wide range of supervised and unsupervised models in corpus-based WSD, including decision lists, decision trees, memory-based learning and bayesian modelling. As these approaches have been thoroughly outlined in detail elsewhere (e.g. Ide & Véronis, 1998; Agirre & Edmonds, 2006b; Navigli, 2009), they are not discussed in detail in the present dissertation.

At an earlier stage of this dissertation (cf. for instance in (Lyse, 2006)) a memory-based approach was pursued, using TiMBL (Daelemans et al., 2007). The memory-based approach was subsequently abandoned because the TiMBL software has been shown to presuppose an initial stage where less informative context features are pruned away prior to training a model (cf. for instance Mihalcea, 2002b). Since our experiments, as described in Chapter (6.4), intend to measure the impact of replacing given context words with Mirrors-derived information about the same words, it was not considered methodologically optimal to use an approach where the $n$ most informative words may be a different set than the set of words that give rise to the $n$ most informative Mirrors-derived context features.

We will now consider the model which is used in the presented experiments, namely Naive Bayes. Several considerations make Naive Bayes suitable for this project's experiments. First, it is vital that we can analyse how and why specific context information influences classification. To this end, the algorithm must be transparent in terms of how training material comes to use at classification time. For instance, a 'black box' neural network is useless for our needs, since is does not allow us to know exactly what has been learnt prior to classification. The Naive Bayes approach, on the other hand, is appreciably simple, in that the training 'model' simply amounts to storing all context information together with frequency counts of occurrence with each sense (see the formal definitions in Section (6.5.2)). The simplicity of Naive Bayes modelling is particularly desirable since we need context information to be given to the system without prior selections (for instance we do not wish to discard less informative context features prior to training). This is important since we wish to compare the direct effect of replacing a context word by its semantic-features (or by the denotation of these semantic-features) according to the Mirrors method. At classification time, we can easily extract information about the contribution of each specific piece of context information when casting a classification vote.

Second, Pedersen (2000) observes that in spite of its simple assumptions Naive Bayes "proves to be among the most accurate techniques in comparative studies of corpus-based word sense disambiguation methodologies" (p. 1) (cf. also

Mooney (1996), Ng (1997a), Leacock et al. (1998), *inter alia*). Since it is a relatively simple and well-understood model with good merits in WSD, it appears well-motivated to apply this methodology for our experiments. Third, Agirre and Stevenson (2006, p. 233-234) compare experimental results conducted by several researchers and observe that Naive Bayes tends to work particularly well for so-called topical features (see Section (3.5.2)), which will be our focal kind of information in this thesis.

### 6.5.2 A formal definition of Naive Bayes

Naive Bayes (NB) is a probability-based greedy model where each possible output is ranked according to probability given the input. For WSD, this may be reformulated as follows: NB estimates the conditional probability $P$ of each sense $s_i$ in a sense inventory $\{s_1, \cdots, s_n\}$, given a set of contextual features $C = \{c_1, \cdots, c_n\}$. Formally, we denote this as

$$P(s_i \mid C), \text{ or } P(s_i \mid \{c_1, \cdots, c_n\})^5. \tag{6.1}$$

The Naive Bayes algorithm computing the probability per TW sense is given in Equation (6.2), and we then choose the sense with the highest probability (the *arg max*):

$$\text{arg max } P(s_i \mid C) = \text{arg max } P(s_i) \prod_{j=1}^{n} P(c_j|s_i) \tag{6.2}$$

There are two parameters to be estimated in this Naive Bayes model: the *a priori* probability of each sense, $P(s)$, and the conditional probability of each context feature $c_j$ given each possible sense, $P(c_j|s)$. Each of these, respectively, are estimated from the corpus material using the equations below.

$$P(s) = \frac{N_s}{N_{tw}} \tag{6.3}$$

$$P(c|s) = \frac{N_{c,s}}{N_s} \tag{6.4}$$

Sense probabilities (Equation (6.3)) are computed as a simple fraction where the numerator is the number of times the given sense was seen in the training material and the denominator is the total number of sense-tagged instances. In our case, since the final training material is a subset $m$ containing only those instances that are tagged with sufficiently frequent senses, we let $N_{tw} = |m|$. The

[5] the formula reads as "the probability of (sense) $s_i$ given (context) $C$"

MLE (maximum likelihood estimation) in Equation (6.4) is computed by counting the number of times that a context feature $c$ correlates with the TW sense $s$, divided by the number of times this sense occurred totally in the training material.

Significantly, NB makes the assumption that the contextual features are conditionally independent given the class. This means that rather than computing the joint probability of $C$ using interdependencies between the features, Naive Bayes assumes that we may instead consider the probability of each contextual feature $c_j$ independently of each other. The NB ranking of each sense given $C$ is based on those contextual features that are known from training. For unknown context words we have no estimation, consequently they are discarded.

Computing the joint product of the independent probability of (known) $c$-features is a sensible approach because the direct probability estimation for $C$ would mean that we must count how many times the exact feature combination in $C$ occurs. Unless we have an extremely simple context model (such as bigrams) the model estimations would then be dominated by zero counts. Individual contextual features $c_i$, on the other hand, may occur in various contexts and we can therefore more easily obtain reasonable statistics for them.

Consider an example where a target word TW occurs $200$ times in our data set. Let us say that it has two senses, $s_1$ and $s_2$, which occur $120$ and $80$ times, respectively. Thus, the probability of each sense is computed as:

$$P(s_1) = \tfrac{120}{200} = .60$$

$$P(s_2) = \tfrac{80}{200} = .40$$

Consider two context words $c_1$ and $c_2$, with the following frequency distributions:
$c_1$ occurs with $s_1$ $140$ times and with $s_2$ $90$ times
$c_2$ occurs with $s_1$ $2$ times and with $s_2$ $10$ times

Using MLE, the $P(c|s)$ for these is then computed as:

$$P(c_1|s_1) = \tfrac{140}{120} = 1.17 \text{ and } P(c_1|s_2) = \tfrac{90}{80} = 1.13$$

$$P(c_2|s_1) = \tfrac{2}{120} = 0.02 \text{ and } P(c_2|s_2) = \tfrac{10}{80} = 0.13$$

Consider now a test situation where the context $C$ only contains the set $\{c_1, c_2\}$. The probability of each sense would then be computed as follows, leaving sense $s_1$ as the most probable class, given $C$.

$$P(s_1|\{c_1c_2\}) = P(s_1) * P(c_1|s_1) * P(c_2|s_1) = 0.60 * 1.167 * 0.016 = 0.09$$

$$P(s_2|\{c_1c_2\}) = P(s_2) * P(c_1|s_2) * P(c_2|s_2) = 0.40 * 1.125 * 0.125 = 0.06$$

### 6.5.3 Training a Naive Bayes model

The training phase in Naive Bayes may be summed up by the following steps.

- Context collection: For each TW instance in the training material, collect context.
- Corpus-based frequency counting, filtering and smoothing
- Computation of the model probabilities

#### Context collection

When collecting data from the parallel corpus, some sifting of information is involved. First, punctuation tokens and tokens with an incomplete analysis are pruned away (e.g. if the lemmatiser marked the part of speech or the lexical entry as UNKNOWN). Second, if the lemmatiser listed more than one syntactic interpretation of a word, the parsing chooses the same interpretation that was chosen for word-alignment (cf. the full explanation in Chapter (4.3.2)). Third, if a corpus token is sense-tagged with a BAG-OF-SINGLETONS partition which we want to ignore (cf. Chapter (4.4.3)), its sense is replaced by an empty string in the source files, i.e. it is then treated as untagged.

#### Corpus-based frequency counts, filtering and smoothing

Each context feature is stored in a hash-table that keeps track of the frequency of co-occurrence of each context feature with each relevant sense of the target word (we do not count how often a context lemma occurs together with the insufficiently frequent senses).

Since the frequency, and hence the probability, of many context features will be zero, smoothing is required in order to avoid zero products in the Naive Bayes computation. As observed by (Navigli, 2009, p. 18), one may apply sense-dependent "smoothing", proposed in (Ng, 1997a) and used for instance in (Bakx et al., 2006, p. 42). Sense-dependent smoothing amounts to replacing zero values by the ratio between the relevant sense probability and the total number of training instances, i.e. $\frac{P(s_i)}{N_{tw}}$. But as remarked by Navigli, this approach has the disadvantage that the probabilities sum to more than 1.

It was therefore chosen to use the beta function ($\beta$) with a very low constant ($0.1$) instead, in which smoothing is performed on *all* observed frequencies, not only on the zero values. Consider the example above (p. 127), where $N = 200$ and where $P(s_1) = .60$ and $P(s_2) = .40$. Let us say there is a context word $c_3$ which occurs twice with $s_1$ and never with $s_2$. In that case,

$P(c_3|s_1) = \frac{2+0.1}{2.2} = 0.95$ and

$P(c_3|s_2) = \frac{0+0.1}{2.2} = 0.05$

Table (6.10) exemplifies some frequency counts with the three different kinds of knowledge sources: a 'traditional' WORD (W) model, a SEMANTIC-FEATURE (SF) model and a RELATED-WORDS (REL-W) model. The example uses the Norwegian noun *stemme*N as the target word (TW). Considering only the Mirrors senses of *stemme*N that are sufficiently frequent (having at least $10$ training instances), this TW is ambiguous between 'voice' (*en glad stemme* 'a happy voice') and 'vote' (*en stemme til det radikale partiet* 'a vote to the radical party'). The frequency counts in the figure are based on a random partitioning of the data set of *stemme*N where $70\%$ are used as training material, which yields $334$ training instances of the target word. The context window was set to $\pm 10$, collecting the $\pm 10$ nearest open-class co-occurrences that were sense-tagged (thus showing the direct loss or gain in abstracting from a set of context words to Mirrors information about precisely the same set of words)

The significance of the selected context features in each of the three models is to show how the frequency counts may differ in the transition from unanalysed context lemmas to Mirrors-derived information. The WORDS model shows six lemmas that actually occur in the context of our target word and their resulting frequencies. The verb co-occurrence *klinge* 'resound, sound', would intuitively be a good contextual indicator of the stemmeVOICE sense, but this lemma was only found to co-occur once with *stemme*|VOICE (and never with *stemme*|VOTE). The five other lemmas, that are intuitively related to *klinge*V, are: *bruse* 'resound', *lyde* 'sound', *ring* 'ring, call', *høres* 'sound' and *synge* 'sing'. All of these lemmas co-occurred only with *stemme*|VOICE and never with *stemme*|VOTE. The question is then if the Mirrors method may improve the statistics counts because similar context words can 'support each other'.

Moving on the SEMANTIC-FEATURES model, then, the lemma *klinge*V has four SEMANTIC-FEATURES, listed in the SF model with the found frequencies per SF. These SEMANTIC-FEATUREs linked the following word senses that actually occurred in the context of *stemme*):

| Model | Entry | Frequency per sense of *stemme*N | |
|---|---|---|---|
| | | **VOICE** | **VOTE** |
| WORDS (W) | *klinge*V | 1 | 0 |
| | *bruse*V | 1 | 0 |
| | *lyde*V | 8 | 0 |
| | *ringe*V | 2 | 0 |
| | *høres*V | 6 | 0 |
| | *synge*V | 8 | 0 |
| SEMANTIC-FEATURES (SF) | $[resound1\|klinge1]$ | 2 | 0 |
| | $[ring1\|ringe1]$ | 11 | 0 |
| | $[sound1\|lyde1]$ | 15 | 0 |
| | $[have1\|vaere1]$ | 2376 | 93 |
| RELATED-WORDS (REL-W) | *klinge*1 | 12 | 0 |
| | *bruse*1 | 4 | 0 |
| | *synge*1 | 17 | 0 |
| | *lyde*1 | 503 | 0 |
| | *ringe*1 | 27 | 0 |
| | *høres*1 | 17 | 0 |

Figure 6.10: Three different co-occurrence models for the ambiguous target word *stemme*N: the unanalysed context WORD, its SEMANTIC-FEATURE or the RELATED-WORDS. Frequencies are counted with a co-occurrence window of the $\pm 10$ nearest word co-occurrences that were sense-tagged per open-class.

$[resound1|klinge1]$ unites the context lemmas *bruse*V and *klinge*V (the SF therefore occurred twice);

$[ring1|ringe1]$ unites *ringe*V, *klinge*V and *synge*V (SF frequency=11);

$[sound1|lyde1]$ unites *høres*V and *klinge*V and *lyde*V (SF frequency=15);

$[have1|vaere1]$ co-occurred 2376 times with *stemme*|VOICE and 93 times with VOTE; the number of individual lemmas that gave rise to this semantic features have not been counted.

The SEMANTIC-FEATURES of one of the actually occurring context lemmas receive higher frequencies than the lemma itself, because the SF frequencies depend on the contextual presence of any word senses that have this SF.

Considering, finally, the RELATED-WORDS model, Table (6.10), this model shows which word senses that were registered in the model based on the sense-tagged context lemmas. The first thing to notice is that the two lemmas *bruse*V and *ringe*V are listed with two Mirrors senses in the Mirrors word bases online[6], the other four lemmas listed in the WORDS model only have one sense in the Mirrors word bases. Both lemmas that are ambiguous according to the Mirrors

[6]Mirrors entries can be looked up online from the following URL: http://maximos.aksis.uib.no:8020/cl/sm/wn-entry.xml. The URL was last verified on April 26, 2011..

method only co-occurred with *stemme*N in one of its senses. The second thing to notice is that the senses in the RELATED-WORDS model, being the counterparts to the actually occurring lemmas in the WORDS model, consistently receive a higher frequency than the semantically unanalysed lemmas. For instance the word sense *klinge*1 was seen more often in the context of *stemme*N than the unanalysed lemma *klinge*V (twelve times vs. once) because the contextual presence of the *word sense* is supported by other word senses in its class of RELATED-WORDS, viz. *bruse*1, *ringe*1 and *lyde*1.

#### Computing the model probabilities

Finally, the model probabilities given each context feature in the training material are computed, as described in Equation (6.3) and Equation (6.4). For each context feature $f_i$, we compute $P(f_i|s_j)$, and for each sense $s_j$, we compute its *a priori* probability (P($s_j$)).

### 6.5.4 Classification in Naive Bayes

Given a previously unseen instance to be classified, the most probable class (sense) is computed by Equation (6.2). For test material, we do not assume that the sense of any context words is known, so we only consider the lemma (the lexical entry together with its POS tag). All context features that actually occur in a given context window contribute to the classification in accordance with how many times it occurs—if a context feature is seen three times in the context of a given target word instance, then its MLE contributes three times to the product in the Naive Bayes formula in Equation (6.2) (p. 126).

It may be remarked that this choice was not obvious, and is something that might deserve a closer study in future work. Intuitively, it would appear reasonable that high-frequent words, such as *være* 'be' and *ha* 'have', are the kinds of words that will most often appear more than once within the context window given a target word instance to be classified, let us say that in that case there are *duplicates* of a context feature within one context window. Moreover, it is generally not very likely that such high-frequent words should be statistically significant for a particular word sense, since their high frequency is usually a reflection of their general ability to occur in the context of any word sense, and for that matter, with any target word. If these assumptions hold, it would be natural to think that the best classification effect is obtained if every context feature only contributes once to classification, i.e. that duplicates of a context feature are removed before computing the joint product of probabilities in the Naive Bayes formula. For instance if, in the WORD model, the lemma *ha* 'have' occurs three times in the context of

a given target word instance, its MLE from training is only added to the product once.

The alternative—to let context features contribute as many times as they are actually seen—should mean that potentially, high-frequent words may dominate the classification outcome, which may be unfortunate if we assume that such context words are not expected to be *salient* in the recognition of a word sense. However, initial experiments indicated, a bit surprisingly, that it was this alternative that gave the best classification results, albeit marginally. The reasons for this is something that could be interesting to follow up in later work; unfortunately there was not time to follow up on this in the present work.

The choice not to remove duplicates should imply that one sticks quite rigidly to the Naive Bayes assumption of an independence between features. This is so because *removing duplicates* would entail, in a way, that you sometimes do assume that there is a dependency between context features, since you assume that it is not a coincidence if a context feature is seen more than once, and since this assumption even causes duplicates to be removed. While it would intuitively seem well-motivated to modify the original Naive Bayes assumption in this way, the preliminary results indicated otherwise, however, so the final choice was to include duplicates.

With the context type option WORD (W), classification proceeds as usual with a Naive Bayes classifier: Look up each lemma associated with the test instance in the training material. If it is known from training, its P(fls) is retrieved and added to the final probability product. Unknown test features are simply not considered.

With the context type SEMANTIC-FEATURES (SF), for each context lemma, its list of possible senses is looked up in the Mirrors lexicon, thus retrieving the union of SFs across senses. Then, each SF is looked up in the training material, and SFs that are known from training contribute to classification. If the context word does not exist in the Mirrors word bases, the word simply did not contribute to classification at all.

With the context type RELATED-WORD (REL-W), for each context lemma, its list of possible senses is looked up in the Mirrors lexicon. Each sense of the context lemma is then looked up directly in the classification model, and any sense that is known from training contributes to classification.

If no context features of a test instance are known from training then we have no basis for using naive classification. A common solution is then to apply a back-off strategy, typically choosing the *a priori* most frequent sense (e.g. Mihalcea & Faruque, 2004). Since we are specifically interested in the contribution of contextual features for WSD, however, such instances are simply left untagged. In this way, we can more easily see the loss or gain in information depending on our level of abstraction.

# 6.6 Evaluation

## 6.6.1 Baseline

As a baseline for comparison, we choose the standard measure in WSD experiments, viz. the *Most Frequent Sense* (MFS) (Section (3.4.2)). In the presented experiments in the chapters to follow, the baseline is computed as follows: Given the most frequent sense in the training material (which could be chosen for every test case), how many of the test instances would then be classified correctly? The MFS is computed based on the number of *test* instances that have the same class as the most frequent sense in the training material.

## 6.6.2 Measuring correct classifications

In our experiments, the classifier does not 'back off' to choosing the most frequent sense if no context features are known from training (cf. p. 132), so it may occur that not all instances to be tagged do receive a tag.

Considering the common evaluation metrics in Section (3.4.2), the following measures will be found in the resulting tables of our classification experiments:

**Precision** (abbreviated Pr): the ratio between the number of correct classifications and the number of classifications that were actually made (Equation (3.2)).

**Recall** (abbreviated Re): the ratio between the number of correct classifications and the number of classifications that should be made (Equation (3.3)).

**F-score** (abbreviated F-s): the weighted harmonic mean of precision and recall (Equation (3.4)).

## 6.6.3 Significance testing

### McNemar's and the sign test

We want to estimate the methodological effect of using differing kinds of contextual information. A test of significance is thus applied to pairs of classification results with a classifier $C1$ and another classifier $C2$.

It was chosen to use a non-parametric sign test due to our data type and the low number of observed changes in the experiment pairs, as explained in what follows: Applying classifiers $C1$ and $C2$ on the same data set, a paired comparison may be used, i.e. counting how many instances were classified the same by both classifiers and how many were differently classified. The paired classification outcomes are

summed up in a contingency table of the type seen in Table (6.4), which contains the counts of whether the test instance had the same classification outcome in both experiments (false-false or true-true) or if there was a difference (false-true or true-false). For instance, the $TF$ cell shows the true-false count. The counts are based on the total set of test instances to be tagged. If a classifier had no knowledge basis for disambiguating a given instance, this counts as a 'false'.

Table 6.4: Contingency table counting the outcomes of a pair of experiments on a sample of $N$ items

| | **True** | **False** | |
|---|---|---|---|
| **True** | $TT$ | $FT$ | $TT + FT$ |
| **False** | $TF$ | $FF$ | $TF + FF$ |
| | $TT + TF$ | $FT + FF$ | $N$ |

Because the resulting counts in the contingency table constitutes category data ('true' or 'false'), it was chosen to use a non-parametric test, i.e. a test that does not make any assumptions about a normal distribution of the data (cf. Rowntree, 2000, p. 124 onwards). Pedersen (2000) and Yarowsky and Florian (2002) were followed in using the non-parametric McNemar's test of significance. McNemar's is, however, considered unreliable when the number of changed classification outcomes ($TF$+ $FT$) is lower than $25$.

For those cases where the number of changed classification outcome is below $25$, we therefore resort to the simpler sign test. The sign test follows the basic intuition that if the 'changed' cells differ significantly, we assume a real change between the samples (an experimental effect).

Both tests were implemented in LISP by the author[7], therefore they are documented for the record.

### The sign test

The p-value of the sign test is based on a binomial probability distribution.

Let $N$ be the number of test instances, of which $m$ = $TF$ + $FT$, i.e. the sum of the 'changed' cells[8]. The null hypothesis ($H_0$) of the sign test is that both experiments perform equally well; that is, the numbers of $FT$ (a performance gain) and $TF$ (a performance loss) should then be roughly equal. Formally, the

[7]This was done to facilitate the generation of output files, since also the Naive Bayes is implemented by the author.

[8]The algorithm for computing the sign test (implemented in Lisp by the author) was found at URL: http://www.stat.yale.edu/Courses/1997-98/101/sigtest.htm. The URL was last verified on April 26, 2011.

chances of seeing an improvement then has $p = 0.5$, and the opposite outcome has the probability value $q = 0.5$:

$H_0 : p = q = 0.5$

In order to identify the alternative hypothesis it is necessary to decide whether to use a one-tailed or two-tailed test. Although we may be more interested in a gain in classification performance, we do not disregard the possibility of an experimental difference in the other direction, i.e. a performance loss. It thus appears most appropriate to use a two-tailed test, which means that our alternative hypothesis is that there *is* a difference but we do not specify the anticipated direction of an observed difference:

$H_a : p \neq q$

Let $Y$ be the more frequent difference ($FT$ or $TF$). The p-value in the sign test is the probability that an observed value $X$ is greater than or equal to $Y$, i.e. $p = P(X \geq Y)$, following the binomial distribution for $n = m$ differences and with probability $p = 0.5$. With a two-tailed test, the resulting p-value is doubled, which in practice means that a bigger difference is required for concluding that an observed difference is significant. If the resulting p-value is less than or equal to our chosen alpha level (the commonly chosen $\alpha$ level is $p = 0.05$), the result is deemed to be statistically significant.

We will first consider an example of what we want and then show how the figures are computed. Let us say that the test sample has $40$ target word instances to be disambiguated, and we test two classifiers C1 and C2 on this material. If $FT = 7$ and $TF = 1$, then $m = 8$, so according to the null hypothesis the differences would follow a $B(8, 0.5)$ binomial distribution. The probability of observing 7 or more positive differences, i.e. $P(X \geq 7)$ is computed as $1-P(X \leq 6) = 1 - 0.965 = 0.035$. The two-tailed P-value is $2 * 0.035 = 0.0703$. Since this P-value is higher than the chosen $\alpha$ level of $0.05$, there is no basis for concluding that the seeming positive effect of applying classifier C 2 is statistically significant. By comparison, it may be noted that the one-tailed P-value is $0.035$, which does indicate that we may reject the null hypothesis since $p \leq 0.05$. This example shows that when $m$ is low (as it will often be in our experiments), the two-tailed sign test is quite conservative in judging a difference to be statistically significant.

The probability $P(X \leq x)$ is calculated as the cumulative probability (the sum of probabilities from 0 to $x$) in a binomial distribution of b(x; m, p), where $m$ is the sum of 'changed' observations ($TF$ + $FT$) and where $p = 0.5$:

$$\text{The cumulative probability} : P(X \leq x) = \sum_{i=0}^{x} b(x; m, p) \tag{6.5}$$

The binomial probability of each value of $x$ is given in Equation (6.6):

$$\text{The binomial formula : } P(x) =_m C_x * P^x * (1-P)^{(m-x)} \tag{6.6}$$

where ${}_mC_x$ is the number of combinations of $m$ objects taken $x$ at a time, not regarding the order of the combined elements (so (B A) = (A B)). ${}_mC_x$ is computed as Equation (6.7), in which $n!$ signifies the *factorial* of $n$: $n! = 1 * 2 * \ldots * n$ (e.g. $4! = 1 * 2 * 3 * 4 = 24$).

$${}_nC_x = \frac{n!}{(n-x)! * x!} \tag{6.7}$$

**McNemar's test**

When the sum of the 'changed' cells exceeds $25$, we use the similar non-parametric McNemar's test, which is computed as:

$$\frac{(1-(FT-TF))^2}{FT+TF} \tag{6.8}$$

Based on the upper critical values of a chi-square distribution with one degree of freedom (one, because we have a two-dimensional contingency table)[9], it is judged whether an outcome is statistically significant or not.

## 6.7 Conclusion

A series of controlled experiments is established to evaluate the Mirrors method, by systematically varying the knowledge source to learn from while maintaining the same experimental framework in terms of the *classification algorithm, data sets, lexical sample and sense inventory*.

This chapter has introduced the knowledge sources to be used in the experiments:

- WORDS (W)
- SEMANTIC-FEATURES (SF)
- RELATED-WORDS (REL-W)

[9]The upper critical values were obtained at: URL: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm. The URL was last verified on April 26, 2011.

As illustrated by the frequency counts in Figure (6.10) (p. 130), there could, at least in theory, be a clear gain in abstracting from traditional context WORDS (WS) to Mirrors-derived information (SEMANTIC-FEATURES or RELATED-WORDS). The experimental setup has been presented, and we have outlined and motivated the choice of Naive Bayes as our classification algorithm, as well as discussing the formal evaluation.

The rest of the experimental framework—the lexical sample and the data sets—being previously unknown to the WSD community, will be presented and discussed in the two next chapters: The automated sense-tagging of the ENPC corpus is presented in Chapter (7), whereas Chapter (8) presents the selection of a lexical sample and the preparation of data sets.

# CHAPTER 7

# AUTOMATIC SENSE-TAGGING OF A PARALLEL CORPUS

## 7.1 Introduction

This chapter presents the method for automated sense-tagging of a parallel corpus, based on the Mirrors method as a knowledge source. This sense-tagged material then forms the basis for moving on to the controlled experiments on a lexical sample.

The presented method sense-tags corpus instances with perfect precision with respect to the Mirrors sense partitions; this thesis addresses the *plausibility* of Mirrors-derived senses and relations between senses. The proposed methodology is applicable for any language pair for which word-aligned corpus material exists, and it may then be applied on both language sides. It has been suggested in the literature to map sense-tags from one language side to the other side in aligned texts (e.g. Diab & Resnik, 2002; Pianta & Bentivogli, 2003), with the possible drawback that the sense-tag of a word in language $L1$ does not necessarily fit as an individual sense for the corresponding $L2$ word. In the proposed approach, each word in each language has a unique set of sense partitions from the Mirrors method which are then used to sense-tag instances.

The basic idea is sketched in Section (7.2), which also discusses some inherent limitations of the method. Section (7.3) presents the results of sense-tagging the English-Norwegian Parallel Corpus (ENPC). As we will see, the number of sense-tagged tokens in the ENPC exceeds that of the biggest manually tagged corpus available, SemCor. Section (7.4) presents and discusses the outcome of the automatic sense-tagger, based on corpus data from the ENPC, for a set of English

lemmas that are commonly used in WSD literature.

## 7.2 Basic idea

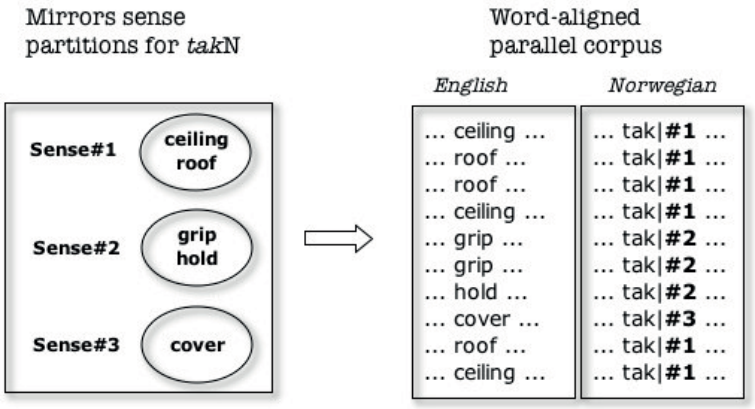



Figure 7.1: Sense-tagging with Mirrors senses: basic idea

The basic idea is illustrated in Figure (7.1) (p. 140), which uses the sense partitions of the Norwegian noun *tak*N as an example[1]: Since the Mirrors sense distinctions are represented by a semantic grouping of the translational correspondents of an ambiguous word into *sense partitions*, we may use the sense partitions as our sense inventory. If an instance of a target word is translated by a member of the first sense partition, the TW instance is tagged with this sense; if an instance corresponds with a member of the second sense partition, then this sense is chosen, etc.

The automatic sense-tagging method is in principle applicable for any language pair for which word-aligned parallel data exist. The method is intrinsically limited by whether the word to be tagged (the *target word*) has an identifiable correspondent in the corresponding sentence. It may be useful to bear in mind that there is a difference between translational correspondents in the linguistic sense as opposed to in the practical sense of automatic word-alignment. The presented sense-tagger implementation requires that a target word instance is actually word-aligned, which does not always happen, for three reasons:

[1]The translational correspondents of this noun were derived from the ENPC manually by Helge Dyvik in the "From Parallel Corpus to Wordnets" project, (Dyvik, 1998).

The first, and linguistically motivated case, is when a corpus instance actually does not have a translational correspondent due to a non-literal rendering of the source sentence on the part of the translator. The two other reasons are practical in nature; in order to be word-aligned, the token must pass lemmatisation, which is not always the case. For instance, foreign lemmas and certain numerical expressions had incomplete analyses after lemmatisation on the Norwegian side, such as *sweatshops* (which occurred once in a Norwegian text) and the numerical expression *13de* 'thirteenth'. The third option is that a token was successfully lemmatised and we would intuitively have identified a word-alignment link, but the automatic word-aligner simply failed to identify an actual correspondent for this particular token. This is typically seen in cases where a sentence in one of the languages corresponds to a much longer or much shorter sentence in the corresponding language.

## 7.3 Automated sense-tagging of the ENPC

### 7.3.1 General

In order to sense-tag the entire ENPC automatically, the computer needed to load the Mirrors word bases while at the same time accessing the corpus. Prior to this dissertation, it had never been attempted to combine the Mirrors word bases and the full corpus material from the ENPC in such a way, and the automatic sense-tagging proved to be slightly more computationally challenging than anticipated.

The automatic sense-tagging of the entire corpus was eventually implemented in LISP by Paul Meurer (Uni Computing). Preceding this, however, a first program code version to sense-tag the ENPC was written by the author. Among other things, these early experiments highlighted the need for a set of heuristic rules to select between multiple lemmas from the lemmatiser as input to the Mirrors method (p. 70). These heuristic rules were implemented by Meurer, who has re-implemented the Oslo-Bergen tagger as well as the Mirrors method code (originating from Dyvik) and who was thus familiar with the computational treatment of both resources. Since the sense-tagger rests on the heuristic selection between lemma alternatives, Meurer simply added a module for automated sense-tagging of the corpus to his existing programming code.

The automatic sense-tagging method was applied to all nouns, verbs and adjectives, and on both language sides of the corpus (English and Norwegian). The last open class, adverbs, is often characterised linguistically as a 'rest class' with members that are mainly held together by syntactic functions. As they do not appear to be particularly semantically interesting in terms of ambiguities and semantic relations, and they have not received much attention in the WSD literature

either, it was decided not to include them in the sense-tagging procedure.

Because the full Mirrors word bases require much computer memory, it was eventually decided to run the tagging procedure only once and to store the result at token level in the XML structure of the ENPC. The automated sense-tagging thus constitutes an extension to the existing ENPC: Word instances in the ENPC are now extended to include an XML tag <sense>, as in Figure (7.2). A sense is then represented numerically as $n/m$, that is, sense $n$ out of $m$ senses. For instance, sense $2/7$ for *company*N in Figure (7.2) corresponds to the sense named *company*2 in the Mirrors bases.

```
<struct type='t-level' id='t_1163_20'>
<feat type='token'>company</feat>
<feat type='position'>20</feat>
<feat type='pos'>N</feat>
<feat type='lemma'>company</feat>
<feat type='sense'>2/7</feat></struct>
```

Figure 7.2: XML representation of a sense-tagged ENPC occurrence of the English noun *company*N

In the following we will present some counts that sum up the results from sense-tagging the corpus. As we will see, the number of sense-tagged tokens exceeds that of the biggest manually tagged corpus available, SemCor.

### 7.3.2 ENPC token counts

Tables (7.1, p. 143) and (7.2, p. 143) sum up the overall number of sense-tagged tokens for the Norwegian and English side of the sense-tagged ENPC, respectively. The counts were obtained by parsing both sides of the ENPC[2]. For comparison, corresponding available statistics about the manually sense-tagged SemCor is included in Table (7.3) (p. 143).

The ENPC token counts in the two first tables sum up, for Norwegian and English, the number of sense-tagged tokens (second row), untagged tokens (third row) and tokens totally (fourth row) per open word class that is considered in this dissertation: nouns, verbs and adjectives. The *coverage* row measures the proportion of sense-tagged tokens, given the total number of tokens per considered word class. Adverb token counts have been included in the table, being an open word class, although they are not sense-tagged in our present material. As opposed to open word classes, the 'Closed-class' line in the tables shows the number of closed-class tokens (for instance prepositions or determiners).

[2]The program code to parse the ENPC and collect statistics about all tokens is written in LISP by the author.

| **ENPC automatically sense-tagged tokens (Norwegian side)** | | | | |
|---|---|---|---|---|
| Word class | sense-tagged | untagged | total | coverage |
| Nouns | 155,567 | 138,291 | 293,858 | .53 |
| Verbs | 145,428 | 94,528 | 239,956 | .61 |
| Adjectives | 45,749 | 66,386 | 112,135 | .41 |
| Adverbs | - | - | 66,992 | - |
| Closed-class | - | - | 552,356 | - |
| **Total** | 346,744 | 299,205 | 1,265,297 | |

Table 7.1: Automatic sense-tagging of the Norwegian side of the ENPC: Tagged and untagged tokens per considered word class. Coverage: the proportion that was tagged.

| **ENPC automatically sense-tagged tokens (English side)** | | | | |
|---|---|---|---|---|
| Word class | sense-tagged | untagged | total | coverage |
| Nouns | 133,742 | 203,393 | 337,135 | .40 |
| Verbs | 145,296 | 107,509 | 252,805 | .57 |
| Adjectives | 43,996 | 55,108 | 99,104 | .44 |
| Adverbs | - | - | 102,569 | - |
| Closed-class | - | - | 548,700 | - |
| **Total** | 323,034 | 366,010 | 1,340,313 | |

Table 7.2: Automatic sense-tagging of the English side of the ENPC: Tagged and untagged tokens per considered word class. Coverage: the proportion that was tagged.

| **SemCor manually sense-tagged tokens (English)** | | | |
|---|---|---|---|
| | sense-tagged | untagged | total |
| **Total** | 234,136 | 302,774 | 676,546 |

Table 7.3: SemCor statistics. From the documentation of SemCor 1.6

These counts disregard punctuation and tokens with an incomplete lemmatisation analysis. Incomplete analyses were typically found with foreign expressions (for instance French *vieille*, occurring once in a Norwegian text and having no lexical entry and no part of speech). Furthermore, the sense-tagged counts disregard so-called bag-of-singleton partitions (explained in Section (4.4.3)): If a token is sense-tagged with such a sense, it counts as *untagged*, since these senses will never be considered when using the data material for Word Sense Disambiguation.

The counts confirm the sum of tokens overall in the corpus (both language sides) found in Johansson et al. (1999/2002)[3]; there are some $2,6$ million tokens when disregarding punctuation (the precise count was $2,606,610$ for both language sides). As Table (7.1) (p. 143) and Table (7.2) (p. 143) show, the coverage figures also seem to confirm the findings of the preliminary experiments in Lyse (2003, 2006): approximately half of the instances are sense-tagged.

The SemCor statistics are retrieved from the documentation of SemCor 1.6[4]. The SemCor documentation includes counts on the number of assigned *senses* per word class, but since a word form in the corpus may be assigned multiple senses, these figures are not directly comparable to the number of sense-tagged ENPC tokens per considered open word class. We therefore only use the total counts from the SemCor documentation. The total word form count was taken from the row *total word forms* in the SemCor documentation. This count includes some more word forms in its *total* counts than the corresponding ENPC counts: The SemCor 'word form' count corresponds to the ENPC 'token' counts in that both exclude punctuation. They differ in that the ENPC token counts do not include foreign words and any other tokens that could not be analysed linguistically by the lemmatiser. In SemCor, foreign words are identified through the attribute *ot* ('other tag') that captures word forms that could not be sense-tagged due to one of a number of alternative cases listed in the documentation, for instance that it is a foreign word or a complex preposition.

The *tagged* count for SemCor was taken from the table row *word forms with semantic pointers*, which includes word forms with one or more WordNet sense-tag, but excludes word forms that are tagged with 'other tags' such as 'metaphor, complexprep' or 'foreignword'. This count seems to be the most compatible to our sense-tagged material: in the ENPC token counts, foreign words are linguistically unanalysed and not even counted in the *total* counts, and metaphors, complex prepositions etc. are not covered in the Mirrors methods (at least not systemat-

[3]URL: http://www.hf.uio.no/ilos/tjenester/kunnskap/sprak/omc/enpc/ENPCmanual.html. The URL was last verified on April 26, 2011.

[4]The documentation of SemCor 1.6 was accessed from URL: http://www.cse.unt.edu/ rada/downloads.html#semcor. The URL was last verified on April 26, 2011., the figures are found in the STATISTICS section of the documentation found at *semcor.htm#sect3*

ically). The *untagged* count was taken from the table row *untagged word forms (cmd=ignore + ot= )*, which is a count of word forms marked as being ignored or tagged with 'other tags' (ot). The untagged counts should thus include word forms from all word classes.

Comparing the amounts of automatically sense-tagged tokens in the ENPC to the manually sense-tagged tokens in SemCor, the results seem promising in that automatic sense-tagger creates a comparably large corpus both for Norwegian and for English, in spite of the face that coverage is relatively modest. The empirical challenge of the present thesis is to investigate how well the Mirrors senses approximate the plausibility of a standardised, manually built lexicon.

### 7.3.3 ENPC lemma counts

According to Zhong and Ng (2009, p. 1), word types (i.e. lemmas) in SemCor have $10$ instances on average.This subsection, too, focuses not on *tokens* (as in the previous section) but on *types*: what is approximately the amounts of data that we can expect when resting on the ENPC and on the Mirrors as our knowledge sources?[5]. All counts are given separately for nouns, verbs and adjectives.

Table (7.4) and Table (7.5) show, first, how many lemmas that occur in the ENPC on the Norwegian and English side, respectively (the first row). Recall that some of these are never recorded in the Mirrors because they are never word-aligned (Section (7.2)). Therefore, the tables also count how many lemmas were registered and not in the Mirrors word bases[6].

It is important to note that bag-of-singleton partitions count as untagged, since one may otherwise wonder why the number of Mirrors lemmas in tables 7.4 and 7.5 are lower than those found in Table (4.1), which counted the number of lemmas in the Mirrors word bases per word class (Chapter (4.1)): Based on the ENPC counts, $6,817$ Norwegian lemmas and $7,529$ English lemmas have only a bag-of-singleton partition, i.e. even though these lemmas exist nominally in the

[5]The program code to parse the ENPC and accumulate statistics for each lemma was implemented in LISP by the author

[6]It may be noted that the counts of ENPC lemmas that are in the Mirrors and that are not in the Mirrors do not add up to the total sum of ENPC lemmas, as they should. In principle, every lemma recorded in the Mirrors word bases should also be sense-tagged at least once, since their presence in the Mirrors method entails that they were word-aligned at least once. It turned out that there is a bug in the code for automatically sense-tagging the ENPC: when a lemma has only one sense partition with only one translational correspondent which is orthographically equal to the target word that gave rise to the sense partitions, it is not sense-tagged. This typically applied to proper names, for instance *Canada* in Norwegian corresponded translationally to *Canada* in English). Since the bug applied to words with relatively low semantic value for our purposes, and since this bug does not apply to the derivation of information in the Mirrors method itself, the programmer was not asked to fix the bug.

**ENPC lemmas included in the Mirrors (Norwegian side)**

| | **Nouns** | **Verbs** | **Adjectives** |
|---|---|---|---|
| ENPC lemmas total | 41,743 | 4,394 | 8,487 |
| ENPC lemmas in the Mirrors | 15,833 | 2,710 | 3,122 |
| Average frequency (+ st.dev.) | $9.83 \pm 45.11$ | $53.66 \pm 646.95$ | $14.65 \pm 70.80$ |
| Maximum | 1705 | 28,096 | 1,817 |
| Minimum | 1 | 1 | 1 |
| Median | 1 | 3 | 2 |
| ENPC lemmas *not* in the Mirrors | 25,892 | 1,683 | 5,365 |
| Average frequency (+ st.dev.) | $2.08 \pm 4.12$ | $2.28 \pm 3.03$ | $2.49 \pm 4.97$ |
| Maximum | 218 | 40 | 195 |
| Minimum | 1 | 1 | 1 |
| Median | 1 | 1 | 1 |

Table 7.4: Statistics on ENPC lemmas (Norwegian side)

Mirrors word bases, they never appear in our present experiments and we therefore also ignore them in the present count of lemmas in the Mirrors word bases.

For the lemmas in the Mirrors as well as for those not in the Mirrors, four simple, descriptive statistics are included to give a rudimentary impression of the number of instances per lemma in each considered open word class: *average frequency* (the mean), maximum value, minimum value and the median. The average value is computed by dividing the sum of each lemma's training corpus size by the number of lemmas. The average value is given together with the standard deviation (*SD*); the $\pm$ value succeeding each average value. SD is a measure of the spread of the averaged values; if the individual observations are close to the mean, SD is low and conversely. Since the SD is consistently higher than the mean value itself, the computed mean cannot be taken to represent the typical training corpus size. Therefore some other simple descriptive statistics are also included: Together with the maximum and minimum number of training instances among the lemmas, the median is included as an indicator of the "typical" training data size. The median is found by sorting all training set sizes from lowest value to highest value and picking the middle one (in case of an even number of observations, the median was computed as the mean of the two middle values).

As can be seen, the average number of sense-tagged instances per lemma ranges between $9.83$ (nouns) and $53.66$ (verbs) on the Norwegian side and between $15.81$ (adjectives) and $54.62$ (verbs) on the English side. Based on the mean value, the average frequency of sense-tagged lemmas is thus comparable to (and actually better than) that of SemCor (Zhong & Ng, 2009, p. 1). If we consider the median, representing the typical number of instances, however, the values are quite low ($3$, $4$ and $2$ for English nouns, verbs and adjectives, respectively, and $1$, $3$ and $2$ for Norwegian nouns, verbs and adjectives, respectively). The tables also show that lemmas not included in the Mirrors word bases are—not surprisingly—

**ENPC lemmas included in the Mirrors (English side)**

| | Nouns | Verbs | Adjectives |
|---|---|---|---|
| ENPC lemmas total | 18,139 | 4,464 | 6,106 |
| ENPC lemmas in the Mirrors | 7,338 | 2660 | 2782 |
| Average frequency (+ st.dev.) | $18.23 \pm 65.42$ | $54.62 \pm 748.60$ | $15.81 \pm 62.97$ |
| Maximum | 1,723 | 34,338 | 1,213 |
| Minimum | 1 | 1 | 1 |
| Median | 3 | 4 | 2 |
| ENPC lemmas *not* in the Mirrors | 10768 | 1804 | 3323 |
| Average frequency (+ st.dev.) | $5.15 \pm 152.62$ | $3.46 \pm 20.95$ | $3.52 \pm 34.52$ |
| Maximum | 15812 | 849 | 1976 |
| Minimum | 1 | 1 | 1 |
| Median | 1.5 | 2 | 1 |

Table 7.5: Statistics on ENPC lemmas (English side).

typically extremely low-frequent.

Tables 7.4 and 7.5 consider the sense-tagged material without considering ambiguity in particular. In the process of identifying a set of target words to be used in a lexical sample (presented and discussed in Chapter (8)), counts were generated on the number of lemmas with *at least a two-way ambiguity*, in order to get a rudimentary impression of the typical data sample sizes for these lemmas. Tables 7.6 and 7.7 provide some descriptive statistics on ambiguous lemmas for Norwegian and English, respectively, using three different values for a parameter defining a *minimum frequency threshold* $MT$. When $MT = 1$ the only requirement is that there must be at least two senses. If $MT = 10$, a lemma is included in the count if it has at least two senses that occur at least ten times each in the corpus. As with the ENPC token counts, bag-of-singleton partitions are ignored when counting the 'senses' of a lemma (these count as untagged).

Considering first all ambiguities, i.e. the ambiguous lemmas where $MT = 1$, the typical training corpus size (the median value) is on the whole not very high (the average value is higher, with a very high standard deviation). The median value for Norwegian nouns, verbs and adjectives, respectively, are $8$, $9$ and $8$; the corresponding medians on the English side are $11$, $8$ and $9$.

The two main reasons for these low median values is that first, ENPC is not a very large corpus. Second, as we have seen, the use of translations as a sense indicator requires that there is an identifiable translational partner, which is not always the case. Given that only approximately half of the corpus is word-aligned (cf. Chapter (4)), and given that lexical units are generally less frequent than function words, the ambiguous lemmas generally do not have large amounts of training instances in our corpus.

Frequency is indeed crucial when employing statistical methods. If one word sense dominates the training corpus whereas another interesting word sense occurs

**Automatically sense-tagged lemmas (Norwegian side)**

| | **Nouns** | **Verbs** | **Adjectives** |
|---|---|---|---|
| Mirrors lemmas ($MT = 1$) | 3,667 | 755 | 974 |
| Average frequency (+ st.dev.) | $32.41 \pm 88.90$ | $141.00 \pm 1214.05$ | $36.93 \pm 121.27$ |
| Maximum | 1705 | 28,096 | 1,817 |
| Minimum | 1 | 2 | 2 |
| Median | 8 | 9 | 8 |
| Mirrors lemmas (MT=3) | 657 | 113 | 151 |
| Average frequency (+ st.dev.) | $62.68 \pm 124.5$ | $122.64 \pm 642.25$ | $81.31 \pm 193.01$ |
| Maximum | 1,371 | 6,785 | 1,658 |
| Minimum | 6 | 6 | 7 |
| Median | 29 | 26 | 28 |
| Mirrors lemmas (MT=10) | 144 | 15 | 31 |
| Average frequency (+ st.dev.) | $106.08 \pm 164.18$ | $87.60 \pm 57.85$ | $153.23 \pm 272.17$ |
| Maximum | 1,125 | 226 | 1,320 |
| Minimum | 21 | 41 | 22 |
| Median | 56 | 62 | 64 |

Table 7.6: Statistics on sense-tagged material in the ENPC (Norwegian side).

**Automatically sense-tagged lemmas (English side)**

| | **Nouns** | **Verbs** | **Adjectives** |
|---|---|---|---|
| Mirrors lemmas ($MT = 1$) | 2987 | 751 | 966 |
| Average frequency (+ st.dev.) | $37.85 \pm 98.19$ | $137.10 \pm 1396.51$ | $35.20 \pm 97.95$ |
| Maximum | 1723 | 34338 | 1,213 |
| Minimum | 1 | 2 | 2 |
| Median | 11 | 8 | 9 |
| Mirrors lemmas (MT=3) | 697 | 105 | 154 |
| Average frequency (+ st.dev.) | $75.80 \pm 151.44$ | $376.36 \pm 3333.88$ | $51.78 \pm 85.93$ |
| Maximum | 1714 | 34328 | 775 |
| Minimum | 6 | 6 | 7 |
| Median | 33 | 18 | 24 |
| Mirrors lemmas (MT=10) | 137 | 9 | 27 |
| Average frequency (+ st.dev.) | $104.14 \pm 103.73$ | $226.0 \pm 525.12$ | $109.48 \pm 88.71$ |
| Maximum | 657 | 1711 | 437 |
| Minimum | 23 | 26 | 21 |
| Median | 66 | 42 | 63 |

Table 7.7: Statistics on sense-tagged material in the ENPC (English side).

but a few times in the entire training corpus, it becomes hard to judge the statistical ability (or lack of such) to "learn" the senses, since extremely low-frequent items give less reliable statistics. It is therefore well-motivated to set a minimum frequency threshold ($MT$) for how many times each word sense must occur in the training material.

A common minimum threshold for statistical WSD is $10$ (cf. for instance Agirre & Martínez, 2004; Zaanen, 2004; Hoste, Hendrickx, Daelemans & Bosch, 2002). This threshold implies that any word sense (in the Mirrors word bases) that occurs less than ten times in the training material is discarded. Formally, all senses of course remain untouched in the Mirrors word base, but in the automatically sense-tagged training material we prune away these low-frequent senses.

Table (7.6) (p. 148) and Table (7.7) (p. 148) include descriptive statistics for ambiguous lemmas where at least two of the senses satisfy a minimum frequency threshold ($MT$) of $10$ and $3$. Since the smallest possible training corpus when MT=$10$ is one where the target lemma only has two senses with ten training instances each, the minimum training corpus size is $20$ (and $6$, in the case of $MT = 3$). All lemmas that satisfy the threshold of MT=$10$ are listed in Appendix 1 (Norwegian ambiguous lemmas) and Appendix 2 (English ambiguous lemmas).

As Table (7.6) (p. 148) and Table (7.7) (p. 148) show, our relatively modest minimum frequency threshold results in a quite dramatic reduction of potential target word lemmas. Consider for instance the Norwegian nouns in Table (7.6), where the counts show that out of $3,667$ ambiguous lemmas only $144$ have at least two senses that occur $\geq 10$.

This amply illustrates one of the strengths of the Mirrors method in the face of sparse data, and potentially one of its setbacks as a source of knowledge for statistical WSD: Since the Mirrors method depends on translational overlap and not on statistics, each observation only needs to be recorded once in order to provide useful information.

Importantly, however, the counts in the above tables do not consider the *plausibility* of the translation-based ambiguity, nor do they consider if the intuitively plausible word senses are sufficiently frequent to be useful for statistical WSD modelling. It may therefore be fruitful to consider some case studies in the next section.

## 7.4 Case studies: some commonly studied English words

As case studies we will consider some of the words that are well-known ambiguous words in the WSD literature (Gale et al., 1992, p. 5, Leacock et al., 1998, p.

4 and 15–16). These lemmas are illustrative examples because their behaviour in our studies shows some successful and some less successful outcomes from the Mirrors method, and the various reasons why.

Table (7.8) lists each lemma and their sense inventory as commonly used in the WSD literature. It should be pointed out that the sense divisions listed in Table (7.8) are not taken to be a final 'gold standard' of sense divisions, but rather a list of helpful indicators of the ambiguities present for each lemma. For instance, Leacock et al. (1998, p. 17) remark that the suggested senses for *work*N are "closely related and therefore difficult for the classifier to distinguish". Furthermore, they state (*ibid.*, p. 4) that sometimes a lemma had senses listed in WordNet that were not included in their experiments because of low corpus frequency.

Table 7.8: English ambiguous words

| **lemma** | **senses** |
|---|---|
| *bank*N | institution vs. land form |
| *bill*N | legal vs. invoice |
| *company*N | business vs. troupe vs. guests |
| *court*N | tribunal vs. sports |
| *drug*N | medication vs. illegal drugs |
| *duty*N | tax vs. obligation |
| *hard*AJ | difficult vs. not soft (metaphoric) vs. not soft (physical)) |
| *land*N | property vs. country |
| *language*N | medium vs. style |
| *line*N | product vs. phone vs. text vs. cord vs. division vs. formation |
| *nail*N | finger nail (body part) vs. metal nail |
| *party*N | political vs. social |
| *position*N | place vs. job |
| *rate*N | monetary vs. freuqency |
| *sentence*N | judicial vs. grammatical |
| *serve*N | supply with food vs. hold and office vs. function as something vs. provide a service |
| *shot*N | sports vs. gunshot vs. opportunity |
| *work*N | activity vs. product |
| *security*N | certificate vs. precaution |
| *stock*N | capital vs. broth |
| *strike*N | work stoppage vs. attack |
| *trade*N | commerce vs. swap |

One could also object that the fact that most of the lemmas in Table (7.8) are polysemous is problematic, since the sense partitions in the Mirrors method are assumed to capture *contrastive* ambiguity. (In the Mirrors method, polysemous senses are expected to emerge through the use of the parameter OverlapThreshold, which varies the granularity within a sense partition; cf. Chapter (4.4.6)). But

the choice of these lemmas is nonetheless well-motivated as we may then get an impression of the practical applicability of the Mirrors method for the purpose of 'classifical' WSD.

The lemmas have been organised into three broad groups according to how they came out with respect to the 'predicted' sense inventory of Table (7.8). The first group contains those lemmas that came out less well in the Mirrors method in that senses were missing compared to the sense inventory predicted in Table (7.8). The second group contains lemmas where we see a similar polysemy between English and Norwegian. The remaining lemmas in the third group are those that appear quite satisfactory in relation to the sense inventory predicted in Table (7.8). The most striking observation in the third group of lemma is that their frequencies are generally very low. They thus illustrate the point that the set-theoretic approach of the Mirrors method (as opposed to a statistical approach) allows the method to generate quite adequate semantic information based on very little information.

**A note on the tables of automatically sense-tagged material and sense distribution**

Following the conventions introduced in Chapter (4.4.2), erroneous 1st $t$-image members (verified manually for all example words in this thesis) are marked by an asterisk.
The English translation in quotes to the left of a Mirrors sense is given by the author to indicate the approximate meaning signalled by the Norwegian sense partition members.
If a sense partition only contains starred entries, i.e. false translational correspondents, the sense partition is marked with a dash to its left, since in that case the meanings of the sense partition members are irrelevant for the target word.
So-called 'bag-of-singleton' partitions (Section (4.4.3)) will be ignored in all WSD experiments of the current thesis, since such partitions do not constitute proper sense partitions. Since they exist in the Mirrors word bases, however, we keep the sense number visible when presenting individual lemmas in order to avoid doubts as to whether a sense is accidentally missing from the table. The sense partition is then replaced by the dummy label BAG-OF-SINGLETONS and it is listed last (cf. *sense*1 of the *shot*N, Table (7.23) (p. 159)).

### 7.4.1 Translational gaps in the corpus

**(*drug*N, *nail*N, *stock*N, *strike*N)**

As was remarked in Chapter (5.2.1) (p. 92) it is difficult to compare sense inventories from different lexical resources, since the absence of some information in one of the resources does not in itself lead us to conclude anything about the quality of what *is* present.

The given selection of case studies illustrate this point: In common for all the lemmas in the present subsection is that some sense is missing in comparison to the sense distinctions predicted in Table (7.8) (p. 150). The missing senses in the current section clearly pertain to accidental gaps in the translational input to the Mirrors method, in virtue of the fact that neither of the Norwegian correspondents of the two lemmas in question could refer to the missing meanings, i.e. the missing senses are not due to a similar polysemy between Norwegian and English. But as we will see with *bill*N in the third category, the Mirrors method may also identify senses that are not expected given the list of predicted senses.

Beginning the lemma *drug*N, where we would expect to see a distinction between 'medication' and 'illegal drugs', the lemma only had one Mirrors sense denoting 'medication', cf. Table (7.9). As for *nail*N, where we expected a distinction between a 'finger nail (body part)' as opposed to a 'metal nail', all its Mirrors senses seem to relate to the 'body part' meaning. As can be seen, *drug*N was only sense-tagged 5 times totally and *nail*N 33 times, which might explain why some translational correspondents were missing.

Table 7.9: *drug*N automatically sense-tagged material and sense distribution.

| **drugN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 5/99 | .05 |
| | Sense distribution: | | | |
| 'medication' | **Sense1:** | {*legemiddel medisin*} ('medication') | 5/5 | 1.00 |
| — | **Sense2:** | BAG-OF-SINGLETONS | — | — |

Table 7.10: *nail*N automatically sense-tagged material and sense distribution.

| **nailN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 33/58 | .57 |
| | Sense distribution: | | | |
| 'claw' | **Sense1:** | {*klo*} | 1/33 | .03 |
| — | **Sense2:** | {*list**} | 1/33 | .03 |
| 'spike' | **Sense3:** | {*nagle*} | 2/33 | .06 |
| 'body part' | **Sense4:** | {*negl*} | 28/33 | .85 |
| 'toe nail' | **Sense6:** | {*tånegl*} | 1/33 | .03 |
| — | **Sense5:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

The lemma *stock*N has a wide range of meanings according to common dictionaries, although only two were listed in the experiments of Leacock et al. (1998); 'capital' and 'broth'. According to the sense-tagged Mirrors material, this lemma was only word-aligned thrice in the entire corpus (disregarding BAG-OF-SINGLETONS partitions) and has three hapax partitions (of which only the translational correspondent in sense 2 was a correct word alignment, *aksje* 'financial stocks').

Table 7.11: *stock*N automatically sense-tagged material and sense distribution.

**stockN**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 3/44 | .07 |
| | Sense distribution: | | | |
| 'share' | **Sense2:** | {*aksje*} | 1/3 | .33 |
| — | **Sense3:** | {*distrikt**} | 1/3 | .33 |
| — | **Sense4:** | {*strømpe**} | 1/3 | .33 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

A similar picture emerges for *strike*N, where we would expect to find an ambiguity between 'work stoppage' and 'attack' according to the commonly used sense inventory in Table (7.8). This lemma was given four Mirrors senses based on 4 word alignments in all (and, hence, it has 4 sense-tagged instances) (Table (7.11) (p. 153)). Of these, Senses 2 and 4 were found to be correct word-alignments, and they both express the meaning of 'attack'. As for the translational correspondents in senses 3 and 5, however, they are actually erroneous word alignments.

Table 7.12: *strike*N automatically sense-tagged material and sense distribution.

**strikeN**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 4/20 | .20 |
| | Sense distribution: | | | |
| 'slap, smack' | **Sense2:** | {*klapp*} | 1/4 | .25 |
| — | **Sense3:** | {*luke**} | 1/4 | .25 |
| 'slap, blow' | **Sense4:** | {*slag*} | 1/4 | .25 |
| — | **Sense5:** | {*tysker**} | 1/4 | .25 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

### 7.4.2 Similar polysemy across languages

**(*land*N, *position*N, *language*N, *line*N, *work*N, *hard*ADJ, *serve*V, *party*N)**

As was indicated in the beginning of this section, the Mirrors method is not expected to delineate polysemous senses at the level of sense partitions. Some of the

lemmas in Table (7.8) have a quite vague meaning; these are the ones listed in this subsection.

For *land*N, Table (7.8) predicted a distinction between 'property' vs. 'country'. But if intuition was not 'primed' by the senses in Table (7.8), it would perhaps be just as plausible to describe *land*N as a lemma with a variety of senses related to the notion of a 'limited area'. The Mirrors sense 1, containing the Norwegian lemma *land* (which has more or less the same polysemy as its English counterpart), comprises lemmas such as *eiendom* 'property', *grunn* 'property', *jord* 'soil/property', *landområde* 'area'. The rest of the Mirrors senses of *land*N are, for one thing (when disregarding false translational correspondents), hapax-correspondents (occurring only once), and would thus be ignored in a statistics-based WSD setting if this was the ambiguous target word. Second, they are best described as pertaining to 'area'-related concepts, such as *beiteland* 'pasture' (sense 3) and *fastland* 'mainland, continent' (sense 5).

Table 7.13: *land*N automatically sense-tagged material and sense distribution.

| **landN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 116/511 | .23 |
| | Sense distribution: | | | |
| 'property, pasture, country' | **Sense2:** | {*Kanaan* del eiendom ende* folk* forfader* forhold* grad* grunn jord jordsmonn* kanal* lag* land landareal* landkjenning* landområde landskap li* liv* lov* mark nordmann* område par* plass sted terreng tid* tun utkant ånd**} | 109/116 | .94 |
| — | **Sense3:** | {*beiteland**} | 1/116 | .01 |
| — | **Sense4:** | {*blod**} | 1/116 | .01 |
| 'mainland' | **Sense5:** | {*fastland*} | 1/116 | .01 |
| — | **Sense6:** | {*innlandsstrøk**} | 1/116 | .01 |
| — | **Sense7:** | {*lensmann**} | 1/116 | .01 |
| — | **Sense8:** | {*million**} | 1/116 | .01 |
| —— | **Sense9:** | {*nes**} | 1/116 | .01 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

With the lemma *position*N, the predicted distinction in Table (7.8) between a 'job' and a 'place' appears quite straightforward. In the Mirrors output (Table (7.14)), however, sense 1 has an overwhelming 120 of all 128 sense-tagged instances, and this sense comprises both of the predicted senses. Similarly to *land*N in the previous example, the English lemma *position*N has a Norwegian correspondent with more or less the same polysemous distinction as the English lemma (*posisjon*).

As regards *language*N, Norwegian *språk* and English *language* have more or less the same polysemy, which makes it unsurprising that the most frequent sense partition is sense 7, consisting only of *språk* (Table (7.13) (p. 154)). Similarly, English *line*N intuitively has a very similar polysemy to Norwegian *linje*N, which

Table 7.14: *position*N automatically sense-tagged material and sense distribution.

| **positionN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 128/197 | .65 |
| | Sense distribution: | | | |
| 'location, attitude, office, placement' | **Sense2:** | {*beliggenhet budsjettsituasjon* holdning kontor plass plassering posisjon situasjon statsminister statsråd* status stilling tittel*} | 120/128 | .94 |
| | **Sense3:** | {*orientering* viten**} | 2/128 | .02 |
| — | **Sense4:** | {*Moskva**} | 1/128 | .01 |
| 'configuration' | **Sense5:** | {*konfigurasjon*} | 1/128 | .01 |
| 'location' | **Sense6:** | {*lokalisering*} | 1/128 | .01 |
| — | **Sense7:** | {*monopolstilling**} | 1/128 | .01 |
| 'stand' | **Sense8:** | {*standpunkt*} | 2/128 | .02 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

is probably the reason why the sense partition containing this Norwegian lemma became the major sense partition, sense 2 (Table (7.16) (p. 156)). Among the senses that are captured in sense 2 are *grense* 'limit', *rad* 'row', *rekke* 'row, series', *rute* 'route, schedule' (e.g. a bus line) and *snor* 'cord'.

Table 7.15: *language*N automatically sense-tagged material and sense distribution.

| **languageN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 90/184 | .49 |
| | Sense distribution: | | | |
| — | **Sense2:** | {*Stortinget**} | 1/90 | .01 |
| — | **Sense3:** | {*fader**} | 1/90 | .01 |
| — | **Sense4:** | {*innslag**} | 1/90 | .01 |
| — | **Sense5:** | {*magasin**} | 1/90 | .01 |
| 'language, tongue' | **Sense6:** | {*mål*} | 3/90 | .03 |
| 'language' | **Sense7:** | {*språk*} | 82/90 | .91 |
| — | **Sense8:** | {*tillegg**} | 1/90 | .01 |
| | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

As for the noun *work*N, the mirrors method generated 9 senses (and one bag-of-singletons partition) but 94% of all its 347 are associated to sense 1, which captures various Norwegian correspondents such as *arbeid* 'labour', *gjerning* 'act, deed', *oppdrag* 'assignment', *oppgave* 'task', *verk* 'creation' (Table (7.16) (p. 156)).

The semantic potential of the Norwegian adjective *hard*AJ is also similar to the English adjective, therefore it is not surprising that most of the translational correspondents of this lemma were grouped into one sense.

*serve*V also ended up with one major sense partition whereas the second sense partition is a hapax partition (Table (7.19) (p. 157)). The second partitions only contains the lemma *ekspedere* 'attend to', which is intuitively related to certain members of the first sense partition (e.g. *betjene* 'wait on' and *servere* 'wait on').

Table 7.16: *line*N automatically sense-tagged material and sense distribution.

**lineN**

| | Total sense-tagged: | | 137/380 | .36 |
|---|---|---|---|---|
| | Sense distribution: | | | |
| 'line' | **Sense2:** | {*dame* ende flekk* forbindelse grense hjerne* klokke* kong* linje modell munn mål* rad rekke rute slekt snor spor stilling* strek stripe stripemønster tank* telefon* tråd vei ønske**} | 109/137 | .80 |
| 'furrow' | **Sense3:** | {*fure rynke*} | 3/137 | .02 |
| 'line' | **Sense4:** | {*Line*} | 1/137 | .01 |
| 'roadway' | **Sense5:** | {*bane*} | 3/137 | .02 |
| — | **Sense6:** | {*brille**} | 1/137 | .01 |
| 'furrow' | **Sense7:** | {*fold*} | 1/137 | .01 |
| 'railway' | **Sense8:** | {*jernbanelinje*} | 1/137 | .01 |
| 'queue' | **Sense9:** | {*kø*} | 8/137 | .06 |
| 'line' | **Sense10:** | {*line*} | 0/137 | .00 |
| 'policy' | **Sense11:** | {*politikk*} | 1/137 | .01 |
| 'speech' | **Sense12:** | {*replikk*} | 2/137 | .01 |
| — | **Sense13:** | {*seil**} | 1/137 | .01 |
| 'utterance' | **Sense14:** | {*setning*} | 1/137 | .01 |
| — | **Sense15:** | {*teglrør**} | 1/137 | .01 |
| 'telephone' | **Sense16:** | {*tele*} | 1/137 | .01 |
| 'twitch' | **Sense17:** | {*trekning*} | 1/137 | .01 |
| — | **Sense18:** | {*utviklingslinje**} | 1/137 | .01 |
| — | **Sense19:** | {*ør**} | 1/137 | .01 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 7.17: *work*N automatically sense-tagged material and sense distribution.

**workN**

| | Total sense-tagged: | | 347/843 | .41 |
|---|---|---|---|---|
| | Sense distribution: | | | |
| 'work' | **Sense1:** | {*alvor* arbeid arbeidsoppgave ark* bok forening foretak forhold* gjerning hoved-sak* jobb kontor krig* oppdrag oppgave sak* skrift verk virke virksomhet yrke*} | 327/347 | .94 |
| 'work' | **Sense3:** | {*arbeide*} | 13/347 | .04 |
| — | **Sense4:** | {*forslag**} | 1/347 | .00 |
| — | **Sense5:** | {*jakt**} | 1/347 | .00 |
| — | **Sense6:** | {*omverden**} | 1/347 | .00 |
| — | **Sense7:** | {*oppvask**} | 1/347 | .00 |
| — | **Sense8:** | {*redskap**} | 1/347 | .00 |
| — | **Sense9:** | {*teglverk**} | 1/347 | .00 |
| — | **Sense10:** | {*trykkeri**} | 1/347 | .00 |
| — | **Sense2:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 7.18: *hard*AJ automatically sense-tagged material and sense distribution.
**hardAJ**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 218/304 | .72 |
| | Sense distribution: | | | |
| 'serious, firm, strict, difficult' | **Sense1:** | {*alvorlig fast god* hard høy kort rund* skarp sterk stor streng tung vanskelig vond*} | 215/218 | .99 |
| 'persistent' | **Sense3:** | {*iherdig*} | 2/218 | .01 |
| — | **Sense4:** | {*satt**} | 1/218 | .00 |
| — | **Sense2:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

The problem with *serve*V seems to be that the English lemma has a Norwegian counterpart with pretty much the same ambiguity, namely *tjene* (in sense 1), which may be used in the context of 'to function/serve as' as well as in the sense of 'providing a service'.

Table 7.19: *serve*V automatically sense-tagged material and sense distribution.
**serveV**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 89/157 | .57 |
| | Sense distribution: | | | |
| 'breed, operate, wait, utilise' | **Sense1:** | {*bedekke betjene bruke dekke fungere gi gjøre henge* servere sitte skulle sone spurte* tjene utgjøre være*} | 88/89 | .99 |
| 'attend to' | **Sense2:** | {*ekspedere*} | 1/89 | .01 |
| — | **Sense3:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

The same problem is seen with *party*N, where Norwegian *parti*N may be used to denote a political party, a group, part, quantity or a match in relationships (although the 'celebration' sense of *party*N does not immediately correspond to Norwegian *parti*, intuitively).

### 7.4.3 Mirrors sense distinctions with plausible sense distinctions

**(*duty*N, *sentence*N, *stock*N *bill*N, *rate*N, *bank*N, *company*N, *court*N, *security*N, *trade*N)**

The remaining lemmas turned out to be quite satisfactory in terms of sense divisions captured through the sense partitions. The most striking point about these lemmas, when considering the number of sense-tagged instances per sense (and hence, the numbers of word-alignments), is that the frequencies are generally very low. This shows that even with a word-alignment standard which is clearly below

Table 7.20: *party*N automatically sense-tagged material and sense distribution.

| **partyN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 264/577 | .46 |
| | Sense distribution: | | | |
| — | **Sense2:** | {*deltaker* representant**} | 2/264 | .01 |
| 'party (political as well as social)' | **Sense3:** | {*Arbeiderparti Avtalepartene* Partiet arbeiderparti avtalepart* fest flokk gruppe kraft* lag part parti prosent* selskap side*} | 245/264 | .93 |
| — | **Sense4:** | {*Arbeiderpartiet**} | 3/264 | .01 |
| — | **Sense5:** | {*Claire**} | 1/264 | .00 |
| — | **Sense6:** | {*autoritet**} | 1/264 | .00 |
| — | **Sense7:** | {*bondeparti**} | 1/264 | .00 |
| — | **Sense8:** | {*bryllup**} | 1/264 | .00 |
| — | **Sense9:** | {*kommunistparti**} | 3/264 | .01 |
| — | **Sense10:** | {*partihierarki**} | 1/264 | .00 |
| — | **Sense11:** | {*partiprogram**} | 2/264 | .01 |
| — | **Sense12:** | {*plassering**} | 1/264 | .00 |
| — | **Sense13:** | {*røst**} | 1/264 | .00 |
| | **Sense14:** | {*sentrumsparti**} | 1/264 | .00 |
| — | **Sense15:** | {*vare**} | 1/264 | .00 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

that of a human annotator, the Mirrors method can generate quite adequate semantic information.

*duty*N has the 'obligation' meaning in sense 1 and the 'tax' meaning in sense 2, which are also the only two senses that are not hapax instances (Table (7.21)). *sentence*N has the 'judicial' meaning in sense 1 and 3 whereas the 'grammatical' meaning is represented in sense 2 (Table (7.22)).

Table 7.21: *duty*N automatically sense-tagged material and sense distribution.

| **dutyN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 41/95 | .43 |
| | Sense distribution: | | | |
| 'obligation' | **Sense1:** | {*forpliktelse plikt*} | 29/41 | .71 |
| 'customs' | **Sense2:** | {*toll tollsats*} | 4/41 | .10 |
| 'charge' | **Sense3:** | {*avgift*} | 1/41 | .02 |
| 'fate, lot' | **Sense4:** | {*lodd*} | 1/41 | .02 |
| 'task' | **Sense5:** | {*oppgave*} | 3/41 | .07 |
| '(in) service' | **Sense6:** | {*tjeneste*} | 1/41 | .02 |
| '(in) service' | **Sense7:** | {*vakt*} | 2/41 | .05 |

The noun *shot*N was given two Mirrors senses, one denoted by *bilde* 'picture' (sense 2) and one denoted by *skudd* 'gunshot' (sense 3). One could in fact say that the Mirrors captured a nice ambiguity that was not seen in Table (7.8) (p. 150). The 'opportunity' sense is thus missed, in that neither of these Norwegian correspondents could be taken to mean a 'chance'. As with the two former lemmas, this lemma had relatively few word-aligned, and thus sense-tagged, instances (19 times).

Table 7.22: *sentence*N automatically sense-tagged material and sense distribution.

**sentenceN**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 34/71 | .48 |
| | Sense distribution: | | | |
| '(judicial)' | **Sense1:** | {*dom*} | 3/34 | .09 |
| 'grammatical' | **Sense2:** | {*setning*} | 29/34 | .85 |
| 'impose a penalty' | **Sense3:** | {*straff*} | 2/34 | .06 |

Table 7.23: *shot*N automatically sense-tagged material and sense distribution.

**shotN**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 19/48 | .40 |
| | Sense distribution: | | | |
| 'picture' | **Sense2:** | {*bilde*} | 4/19 | .21 |
| 'gunshot' | **Sense3:** | {*skudd*} | 15/19 | .79 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*bill*N misses the 'legal' meaning, but interestingly it has captured the sense of a 'beak' in sense 2, which is not captured in the predicted sense inventory suggested in Table (7.8) (p. 150). The predicted 'invoice' is captured in Sense 3 (Table (7.24)). As for the lemma *rate*N, the Mirrors succeeds in grouping together four plausibly related Norwegian correspondents pointing to the predicted sense of 'pace, rhythm' in sense 2. The 'monetary' sense occurs in senses 3, 4, 5 and 9, although not all of these are found to be equally frequent (Table (7.25)).

Table 7.24: *bill*N automatically sense-tagged material and sense distribution.

**billN**

| | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 35/62 | .56 |
| | Sense distribution: | | | |
| — | **Sense1:** | {*fugl**} | 1/35 | .03 |
| 'beak' | **Sense2:** | {*nebb*} | 12/35 | .34 |
| 'invoice, check' | **Sense3:** | {*regning*} | 22/35 | .63 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

The lemma *bank*N was word-aligned and sense-tagged relatively rarely, but still has 48% of its instances associated to the meaning of a 'river bank' (sense3) whereas sense 9 points to the 'monetary' sense (Table (7.26)).

The lemma *company*N, with a predicted ambiguity between 'business', 'troupe' and 'guests' was found to have a relatively nice ambiguity in senses 2, 3, 4 and 6 (Table (7.27)).

Considering *court*N, senses 1 and 2 encapsulate the 'judicial' senses whereas sense 3 denotes the 'royal' sense. Also, sense 5 has identified the meaning of courting a potential partner (Table (7.28)).

Considering the lemma *security*N, the author is not entirely sure what the 'certificate' sense would contain and not. The Mirrors method generated (in sense 2) the notion of 'safety', whereas sense 3 denotes the sense of 'public security'

Table 7.25: *rate*N automatically sense-tagged material and sense distribution.

**rateN**

| | Total sense-tagged: | | 21/154 | .14 |
|---|---|---|---|---|
| | Sense distribution: | | | |
| 'rhythm' | **Sense2:** | {*fart rytme takt tempo*} | 6/21 | .29 |
| 'amount' | **Sense3:** | {*beløp tilfelle*} | 2/21 | .10 |
| 'interest rate' | **Sense4:** | {*rentenivå rentesats*} | 3/21 | .14 |
| 'charge' | **Sense5:** | {*avgift*} | 1/21 | .05 |
| | **Sense6:** | {*forbrytelse**} | 1/21 | .05 |
| 'degree' | **Sense7:** | {*grad*} | 1/21 | .05 |
| — | **Sense8:** | {*kur**} | 1/21 | .05 |
| 'price' | **Sense9:** | {*pris*} | 3/21 | .14 |
| — | **Sense10:** | {*vekstrate**} | 3/21 | .14 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

∗Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 7.26: *bank*N automatically sense-tagged material and sense distribution.

**bankN**

| | Total sense-tagged: | | 27/202 | .13 |
|---|---|---|---|---|
| | Sense distribution: | | | |
| 'monetary' | **Sense2:** | {*bank barnepleier**} | 1/27 | .04 |
| 'riverside' | **Sense3:** | {*barn* bredd kant land rad*} | 13/27 | .48 |
| — | **Sense4:** | {*avdeling**} | 1/27 | .04 |
| — | **Sense5:** | {*bankdirektør**} | 1/27 | .04 |
| — | **Sense6:** | {*bankkonto**} | 1/27 | .04 |
| 'flower bed' | **Sense7:** | {*bed*} | 1/27 | .04 |
| — | **Sense8:** | {*sandbanke**} | 1/27 | .04 |
| — | **Sense9:** | {*sentralbank**} | 6/27 | .22 |
| — | **Sense10:** | {*skog**} | 1/27 | .04 |
| 'riverside' | **Sense11:** | {*skråning*} | 1/27 | .04 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

∗Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 7.27: *company*N automatically sense-tagged material and sense distribution.

**companyN**

| | Total sense-tagged: | | 85/183 | .46 |
|---|---|---|---|---|
| | Sense distribution: | | | |
| 'enterprise' | **Sense2:** | {*bedrift firma*} | 25/85 | .29 |
| 'assembly, branch' | **Sense3:** | {*avdeling forsamling følge mann*} | 8/85 | .09 |
| 'party' | **Sense4:** | {*lag selskap*} | 44/85 | .52 |
| — | **Sense5:** | {*Frans**} | 1/85 | .01 |
| 'unit' | **Sense6:** | {*kompani*} | 6/85 | .07 |
| — | **Sense7:** | {*transport**} | 1/85 | .01 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

∗Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 7.28: *court*N automatically sense-tagged material and sense distribution.

| **courtN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 72/152 | .47 |
| | Sense distribution: | | | |
| 'law' | **Sense1:** | {*For* rett*} | 40/72 | .56 |
| 'law' | **Sense2:** | {*domstol*} | 11/72 | .15 |
| 'royal' | **Sense3:** | {*hoff*} | 18/72 | .25 |
| — | **Sense4:** | {*kunst**} | 1/72 | .01 |
| 'flirt' | **Sense5:** | {*kur*} | 2/72 | .03 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

through social services and benefits.

Table 7.29: *security*N automatically sense-tagged material and sense distribution.

| **securityN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 60/115 | .52 |
| | Sense distribution: | | | |
| 'safety' | **Sense2:** | {*sikkerhet trygghet*} | 57/60 | .95 |
| 'social security' | **Sense3:** | {*trygd ytelse*} | 2/60 | .03 |
| — | **Sense4:** | {*vei**} | 1/60 | .02 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Finally, *trade*N has grouped quite neatly a group of translational correspondents that plausibly belong together. In sense 2 there are words pertaining to *bransje, handel* 'commerce', and the same area of meaning is covered in sense 3. Then, sense4 (occurring only once) has recorded the sense of 'swap'.

Table 7.30: *trade*N automatically sense-tagged material and sense distribution.

| **tradeN** | | | | |
|---|---|---|---|---|
| | Total sense-tagged: | | 67/153 | .44 |
| | Sense distribution: | | | |
| 'commerce' | **Sense2:** | {*bransje fag handel næringsliv omsetning yrke*} | 56/67 | .84 |
| 'commerce' | **Sense3:** | {*Trondheim* gruppe* landbruksvare* område verden**} | 5/67 | .07 |
| 'swap' | **Sense4:** | {*bytte*} | 1/67 | .01 |
| — | **Sense5:** | {*frihandel**} | 1/67 | .01 |
| 'commerce' | **Sense6:** | {*håndverk*} | 2/67 | .03 |
| | **Sense7:** | {*utland**} | 1/67 | .01 |
| 'commerce' | **Sense8:** | {*verksted*} | 1/67 | .01 |
| — | **Sense1:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

## 7.5 Conclusion

In this chapter the method for automated sense-tagging of an entire corpus has been presented. The sense-tagging of the ENPC, constituting our material for the now succeeding WSD experiments, has been quantified in terms of how many tokens could be tagged and in terms of counts on the typical training material size.

We have seen that one of the strengths of the Mirrors method in the face of sparse data is that it depends on translational overlap and not on statistics. Each observation thus only needs to be recorded once in order to provide useful information, but this is at the same time potentially one of its setbacks as a source of knowledge for statistical WSD. Based on earlier work resting on manually extracted translational correspondences from the ENPC (Lyse, 2003), the possibility of 'enlarging' the information value of small corpus data sets by adding Mirrors-derived paradigmatic dimension certainly could offer an interesting trajectory.

Also, the presented figures of this chapter show that the present sense-tagged corpus is larger than the biggest hand-crafted text corpus with sense-tags. This promising fact comes with a caveat: the usefulness of such amounts of tagged data depends on the *plausibility* of the Mirrors method. So the main question, to be pursued in the following chapters, is: what about the plausibility of the sense-tagged material, and what is the practical applicability of these resources for WSD?

# CHAPTER 8

# AN EXPERIMENTAL LEXICAL SAMPLE

## 8.1 Introduction

The present thesis aims for a lexical sample evaluation. During a panel at SENSEVAL-3 it was stated that the lexical sample task is becoming less and less interesting, as the disambiguation of a single target word in a sentence is not useful in most human language technology applications. But this argument is mainly based on practical considerations of scalability. WSD experiments that rest on manual labour may be feasible only for a small set of words, and in that case one may argue that the *approach* will not scale up, but it seems obvious that the experimental environment (the lexical sample task) is then not, by itself, part of the scalability problem.

Another reason for a certain scepticism against the lexical sample approach is the acknowledgement of WSD as a 'word expert' task. With a small lexical sample one cannot ascertain whether the observed results will apply to other words in an all-words setting. But the major advantage of a lexical sample is that it is feasible to analyse the results in some detail. Glizzio et al. (2005, p. 1) argue that originally, the lexical sample task was intended to provide a clearly defined framework for experiments: "We think that a lexical sample WSD should regain its original explorative role and possibly use a minimal amount of training data, exploiting instead external knowledge acquired in an unsupervised way to reach the actual state-of-the-art performance".

Indeed, several recent research projects seem to adopt this thinking. For instance, Specia et al. (2009) explore the contribution of various deep and shal-

low various knowledge sources for WSD. They study the behaviour of different knowledge sources, evaluating on data sets from the SEMEVAL-2007 English lexical sample task, using corpus examples of 65 verbs and 35 nouns. Their data sets have an average of 222 examples for training and 49 for testing per target word (the minimum no. of training instances is 19 training examples and 2 for test material, the maximum is 2,536 training instances and 541 test instances). The Swedish lexical sample introduced in SENSEVAL-2 contained 40 lemmas (20 nouns, 15 verbs and 5 adjectives), totalling 8718 training/development instances and 1525 test instances, which on average yields 218 training instances and 38 test instances per word (Lager & Zinovjeva, 2001).

For the present dissertation a lexical sample is well-motivated because the Mirrors method, being an experimental knowledge resource, makes it particularly desirable to focus on a tractable lexical sample to obtain a good analysis of the behaviour of the classifiers. Norwegian being the mother tongue of the author, it was chosen to use Norwegian target words to facilitate the experiment analyses.

Since this dissertation introduces a set of target words with a sense inventory that is not commonly known in the WSD community, the target words and data sets are documented in some detail. In brief, the current dissertation casts 15 target words (10 nouns, 3 adjectives and 2 verbs). The data set is compiled from the ENPC, consisting of all instances that were sense-tagged automatically and manually sense-tagged instances that could not be sense-tagged automatically. By combining automatically and manually sense-tagged corpus instances we acquire a larger material from the ENPC while maintaining the opportunity to use Mirrors-derived information about context words, since the ENPC is automatically sense-tagged. This is in line with the stated experimental framework of Chapter (6), in which the automatic sense-tagging of Chapter (7) is seen as a separate task from that of systematic experiments with context information, which we embark on now. The total data set has on average 269 $\pm$337 corpus instances, the minimum number being 54 instances and the maximum being 1324 (based on the counts in Table (8.19) (p. 193)).

Section (8.2) states the criteria for selecting a lexical sample. Section (8.3) presents an overview of how the data sets are produced. Section (8.4) presents the 15 target words, comparing the Mirrors sense inventory for each word against that of a common Norwegian dictionary. Having presented the lexical sample, Section (8.5) then discusses issues related to the development of the data sets, both regarding the quality of the automatically sense-tagged data and regarding the manual sense-tagging of the instances that were not sense-tagged automatically. Finally, Section (8.6) discusses the representativity of the data sets.

## 8.2 Criteria for selecting a lexical sample

When choosing target words for WSD classification, we do not consider the semantic relatives of a word; the pertinent questions concern the following two considerations:

1. Ambiguity
   An appropriate target word lemma must have at least two plausible senses in the Mirrors base, and satisfy a minimum frequency threshold of occurrences per such sense (senses below this threshold are discarded).

2. Total frequency
   Target words with a high total number of automatically sense-tagged TW instances are preferred to those with a lower total frequency.

The first criterion states that for the target words in our lexical sample, it must be plausible to expect that their sense divisions from the Mirrors method correlate, at least to some extent, with those that we would expect to find in a common dictionary. That is, we would like the sense inventory of our initial target words to be as intuitively uncontroversial as possible. (As we will see, however, this does not mean that the senses of our selected target words are unproblematic.) One might immediately object, 'but is not that cheating, if part of the question in this dissertation is whether the Mirrors generate senses that are plausible in such a way that we expect them to have contextual correlates?' There are two principal reasons for answering 'no' to this.

At the heart of both reasons is the crucial point that our controlled experiments aim to test the Mirrors-derived knowledge about *context* words, and not the target words in themselves. First, the kinds of contextual knowledge are varied by starting from context words and by tentatively adding Mirrors-derived knowledge about the observed context words. It therefore simply seems wise to begin from a selection of words with sense distinctions that are as uncontroversial as possible, i.e. that are as close to a classical WSD experiment as possible. In this way, we provide a framework for focussing, not on how difficult it was (in itself) to learn the senses of the ambiguous target word, but on the effect of varying the *contextual* information that the classifier learns from.

Second, as we will see, the plausibility of word senses is specifically targeted in (Chapter (10)), in which we train classifiers on sense-specific Mirrors-information on the one hand and, on the other hand, on the union of Mirrors-information associated to a context lemma. In this way, the evaluation of sense plausibility is not limited to the learnability of the words in a limited lexical sample.

The second criterion relates to the "data sparseness problem". Given the relatively small corpus resource at our disposal, training material is actually quite limited in size even in the best cases. It was considered a relevant concern to avoid that the target words in the lexical sample are so low-frequent that it becomes hard to justify that the machine learning results say anything informative at all. The lexical sample experiments function as a kind of 'proof of concept', intended to show how well the proposed method may perform, given optimal data from a corpus resource. Hence, it is desirable to begin with those target words where the training material maximally satisfies the general data set desiderata in a corpus-based classification approach to WSD. The frequency threshold is currently set to 10, which is a common minimum threshold for statistical WSD (cf. Agirre & Martínez, 2004; Zaanen, 2004; Hoste, Hendrickx, Daelemans & Bosch, 2002).

## 8.3 Developing the data sets for WSD

**Some basic definitions**

The present thesis builds one development data set and one held-out test set per target word (TW) in the lexical sample.

The development of a machine learning classifier for WSD generally presupposes three data sets: a *training set*, a *development test set* and a set of *previously unseen test instances* (a held-out data set). From the training set the frequencies for estimating the relevant probabilities are retrieved. The development test set is used while developing a final classifier, assessing for instance the most optimal context window size. Then the final classifier should be tested on a second test set (the held-out test set) that was never a part of the development experiments.

In the not uncommon case of sparse sense-tagged data in WSD experimens, training and testing on the same data set through so-called *cross-validation* is often used (e.g. Specia et al., 2009; Pedersen, 2000; Hoste, Hendrickx, Daelemans & Bosch, 2002; Ng, 1997a). $n$-fold cross-validation means that the original training material is randomised and partitioned into $n$ parts, or *folds*. $n$ experiments are then run, each time with a new fold as test data whereas the remaining material serves as training instances. The final result is presented through the averaged outcome of each sub-classifier. (See e.g. Daelemans et al., 2007, p. 10 for more details on cross-validation tests.)

For the development phase, this dissertation, too, rests on cross-validation. It was then chosen to opt for a 5-fold validation, as in Pedersen (2000). Thus, two separate data sets are developed for the presented lexical sample experiments; one development set (where we train and test on the same data set using cross-validation) and one held-out test set (for testing the final classifier in each experi-

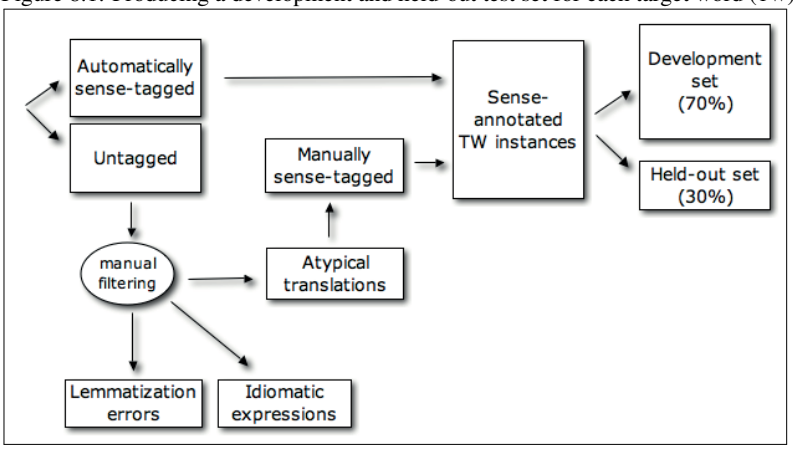
ment).

As regards the final test set, it should ideally be taken from an independent corpus, i.e. from another resource than the ENPC: Since the Mirrors information and the resulting WSD classifiers are all based on the ENPC, the most informative test situation would be one that tests whether the predicted classes of semantically related words would be of any use outside the ENPC. Indeed, as was remarked in Chapter (1), the intention at an earlier stage of this thesis was to test Mirrors-based WSD classifiers on test material from the LOGON project. Regrettably, time has not permitted to do this final test in the way planned. So in the absence of an external test set, we resort to using previously unseen material from the ENPC for validation.

### Developing sets of sense-annotated TW instances

The production of sense-annotated data set is produced as illustrated in Figure (8.1) (p. 167).

Figure 8.1: Producing a development and held-out test set for each target word (TW)




For each instance of the ambiguous word (the *target word*) in the corpus, a program checks if the instance was automatically sense-tagged on the basis of its translation[1]. If the instance is automatically sense-tagged, the instance is added directly to the pool of sense-annotated TW instances (see the upper right of Figure (8.1)).

[1]The program is written in LISP by the author.

The not automatically sense-tagged instances (the 'untagged' instances, for short) are inspected manually by the author. The instances untagged by the automatic sense-tagger were found to belong to three categories, of which two categories represented unwanted material (lemmatisation errors and idiomatic expressions). The final category contains instances that were not word-aligned (and therefore not sense-tagged) because of atypical (non-literal) choices on the part of the translator, i.e. its 'atypicality' pertains to the translational relation and not to the TW instance, seen from a monolingual point of view. Instances in the latter category are sense-tagged manually by the author and added to the set of sense-annotated TW instances. The resulting set of TW instances is then shuffled and $70\%$ of the instances (picked randomly) are allotted for development whereas the remaining $30\%$ are used as the final, held-out test material.

Concerning the manual sense annotation, observations and problems that are related to specific words and senses in the lexical sample are presented and discussed in Section (8.5). At a more general level it may be remarked that although it is often considered methodologically sound to use inter-annotator agreement (e.g. Ide et al., 2002; Ide & Erjavec, 2001), resources were not allocated in this project for that purpose. In retrospect, an interesting extension to the current project might have been be to compare the automatically sense-tagged material against a measure of inter-annotator agreement on the same data sets. This would provide a nice way to evaluate the automatic sense-tagger based on the Mirrors, seen as an alternative to producing sense-annotated material manually.

But whereas an inter-annotator agreement is confined to judgments of sense-tags in context, this thesis aims to judge further aspects of the Mirrors method than the sense-tags. For this reason it was chosen to focus on what a practical evaluation in Word Sense Disambiguation may tell us about the Mirrors method as a knowledge source, since we may then test senses as well as semantic relations between senses. Moreover, the target words (for which data sets for testing are tagged manually) were selected with the intention that they should rest on relatively uncontroversial sense distinctions that also satisfy a minimum frequency threshold in the corpus, based on the automatically sense-tagged material. This was done precisely in order to leave the formal evaluation of Mirrors-derived semantic representations (senses and semantic relations) to be conducted by systematic experiments based on context. Hence, the need to measure inter-annotator agreement on these test sets of the target words was not attributed strong importance.

Table 8.1: 15 initial target words: overview of the Mirrors sense inventory to be used in experiments (only the sufficiently frequent senses)

| | Lemma | Senses |
|---|---|---|
| Nouns | *lag*N | *lag*5 ('layer'<br>*lag*8 ('team') |
| | *fyr*N | *fyr*1 ('guy')<br>*fyr*2 ('fire') |
| | *utvalg*N | *utvalg*1 ('selection')<br>*utvalg*3 ('committee') |
| | *rot*N | *rot*2 ('root')<br>*rot*7 ('mess') |
| | *plan*N | *plan*1 ('scheme')<br>*plan*2 ('level') |
| | *valg*N | *valg*1 ('choice')<br>*valg*2 ('election') |
| | *slag*N | *slag*1 ('kind, type')<br>*slag*2 ('battle')<br>*slag*3 ('blow')<br>*slag*12 ('stroke') |
| | *tak*N | *tak*2 ('grasp')<br>*tak*4 ('ceiling')<br>*tak*7 ('roof') |
| | *stemme*N | *stemme*1 ('voice')<br>*stemme*3 ('vote') |
| | *liv*N | *liv*1 ('life')<br>*liv*12 ('waist') |
| Adjectives | *full*AJ | *full*1 ('complete')<br>*full*2 ('drunk') |
| | *gal*AJ | *gal*1 ('crazy')<br>*gal*2 ('incorrect') |
| | *frisk*AJ | *frisk*1 ('fresh')<br>*frisk*4 ('healthy') |
| Verbs | *trykke*V | *trykke*1 ('squeeze')<br>*trykke*2 ('print') |
| | *utsette*V | *utsette*1 ('expose')<br>*utsette*3 ('postpone') |

## 8.4 Presenting the lexical sample

In this section each target word is briefly presented with respect to its automatically sense-tagged material. Target words were selected on the basis of the automatically sense-tagged material, since the quality of the automatically untagged instances was not known *a priori*. Then we move on in the next section (Section (8.5)) to the manually sense-tagged material.

In the current section the Mirrors sense inventory is compared against the commonly available Norwegian dictionary Bokmålsordboka[2].

[2]URL: http://www.dokpro.uio.no/ordboksoek.html. The URL was last verified on April 26, 2011.

For the benefit of the non-Norwegian reader, the original Bokmålsordboka entry has been simplified such that it only shows the sense enumeration and provides an English account (made by the author) of each sense. Homonymy is denoted by roman numbers (I, II, III) whereas subsenses (polysemy) are given by arabic enumeration (1, 2, 3). Sometimes the dictionary entry is too fine-grained (detailed) to be useful; therefore the simplified dictionary entry is sometimes omitted and its list of subsenses are just summed up in the text.

For the sake of overview, the target words are listed in Table (8.1) (p. 169). Presenting each lemma in the following, a table is given which shows the sense frequencies of each Mirrors sense based on automatic sense-tagging.

The first such table of Mirrors senses is given in Table (8.2) (p. 172). The table lists the total number of sense-tagged instances and the sense distribution. The senses are listed in the same order as their numbering from the Mirrors method, except that senses with a frequency below the minimum threshold of 10 instances are listed after the sufficiently frequent senses. Also, so-called BAG-OF-SINGLETONS senses (cf. Section (4.4.3)) are listed at the bottom. Each number of automatically sense-tagged instances is given as a simple frequency count (the penultimate column) and the corresponding proportional figure (the last column). Starred lemmas are false translational correspondents from automatic word alignment (manually verified by the author for this thesis).

### 8.4.1 Target nouns

#### *lag*N

The noun *lag*N is useful as a target word as there is a quite clear intuitive ambiguity between its Mirrors senses of ‘team’ and ‘layer’.

The common Norwegian dictionary Bokmålsordboka[3] defines only one main sense with 12 subsenses.

[3]URL: http://www.dokpro.uio.no/ordboksoek.html. The URL was last verified on April 26, 2011.

***lag*N senses based on Bokmålsordboka**

I. 1 layer
2 a social class (*de høyere sosiale lag* 'the upper social strata')
3 something grouped together (e.g. *nabolag* 'a neghbourhood')
4 social companionship (*i godt lag* 'in good company')
5 party, company (*50-års**lag*** '50 year anniversary')
6 club, team (*spille på et lag* 'play on a team')
7 group/organisation that works together
8 (fixed expression) *stå ved lag* 'remain in force',
9 (fixed expression) a property (of personality), e.g. *hjertelag* 'compassion'
10 (fixed?) an ability or a way (*ha godt lag med barn* 'have a way with children'
11 (fixed) mood, with genitival after the preposition *til* 'to': *gjøre noen til lags* 'make someone content'
12 (fixed) *gi noen det glatte lag* 'yell at someone'

Bokmålsordboka's subsense 1 points to a 'layer' sense (*et tynt lag* 'a thin layer'), which corresponds quite nicely with the Mirrors sense 5. The dictionary subsense 2 refers to a social class (*de høyere sosiale lag* 'the upper social strata', which is similar both to the Mirrors senses 5 ('layer') and 1 ('class, company, party'), although the Mirrors sense 1 is not sufficiently frequent to be included in our training material. Indeed, an inspection of the automatically sense-tagged instances revealed that the Norwegian collocation *sosiale lag* 'social layer/stratum/class' appeared twice with the Mirrors sense 1 ('company, class'), and once with sense 5 ('layer'), suggesting that these two senses are related.

The dictionary subsenses 3, 4, 5 comprise notions of a social group, such as *nabolag* 'a neghbourhood' or *i godt lag* 'in good company', which are perhaps closer to the Mirrors sense 1, if any. Bokmålsordboka's subsenses 6 and 7 encompass the same meaning as Mirrors sense 8, 'team': a club, team or organisation (e.g. *spille på lag* 'play on the team'). Finally, Bokmålsordboka's subsenses 8, 9, 10, 11 and 12 comprise various fixed and metaphoric expressions where it is difficult to find an appropriate literal translation of *lag* alone (e.g. *stå ved lag* 'remain in force', *hjertelag* 'compassion', *ha godt lag med barn* 'have a way with children', *gjøre noen til lags* 'make someone happy', *gi noen det glatte lag* 'let someone have it'.

The Mirrors senses generated eight senses of *lag*N, of which one is a so-called BAG-OF-SINGLETONS sense (cf. Section (4.4.3)) and of which only two are suffi-

| **lagN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 47/92 | .51 |
| Sense distribution: | | | |
| **Sense5:** | {*layer*} | 23/47 | .49 |
| **Sense8:** | {*team*} | 13/47 | .28 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense1:** | {*class company party*} | 6/47 | .13 |
| **Sense2:** | {*land* way*} | 2/47 | .04 |
| **Sense3:** | {*air*} | 1/47 | .02 |
| **Sense4:** | {*lawyer**} | 1/47 | .02 |
| **Sense6:** | {*level*} | 1/47 | .02 |
| **Sense7:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.2: Mirrors entry: *lag*N automatically sense-tagged material and sense distribution.

ciently frequent, namely sense 5 ('layer') and sense 8 ('team'). The number of training instances is not high, only 47 out of 92 (51%) instances of *lag*N were automatically sense-tagged. Considering only the sufficiently frequent senses, the Mirrors sense 5 occurred 23 times and sense 8 occurred 13 times.

### *fyr*N

This noun is interesting because it has a very clear contrastive ambiguity. Bokmålsordboka divides it into three contrastive senses:

***fyr*N senses according to Bokmålsordboka**

I. guy, type (*en hyggelig fyr* 'a nice guy')

II. fire, light, furnace (*sette fyr på* 'set fire on'; *har du fyr?* 'have you got a light?')

III. lighthouse, beacon

The Mirrors method generated three senses that encapsulate the same senses as those predicted in Bokmålsordboka, although only the two first senses are sufficiently frequent. The 'lighthouse' sense (the Mirrors sense 3) is thus lost in the training material for this noun.

| **fyrN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 50/84 | .60 |
| Sense distribution: | | | |
| **Sense1:** | {*chap fellow guy man type*} | 36/50 | .72 |
| **Sense2:** | {*fire*} | 12/50 | .24 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense3:** | {*lighthouse*} | 2/50 | .04 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.3: Mirrors entry: *fyr*N automatically sense-tagged material and sense distribution.

The number of instances per sense is on the whole not high, which makes *fyr*N interesting as a test case to see if there is any gain in abstracting from context words to classes of semantically related context words.

### *utvalg*N

According to Bokmålsordboka, this polysemous noun has three related senses:

| ***utvalg*N senses according to Bokmålsordboka** |
|---|
| I. 1 a selection, variety (*et utvalg av sanger* 'a selection of songs') |
| 2 a range, an amount between which one may select (*et dårlig utvalg av bøker* 'a poor range of books') |
| 3 committee (*oppnevne et utvalg* 'appoint a committee') |

Considering the two sufficiently frequent senses in the Mirrors method (Table (8.4) (p. 173)), the Mirrors sense 1 seems to encompass the two first-mentioned senses in Bokmålsordboka; denoting both a 'selection' as well as 'range'. The Mirrors sense 2 encapsulates the sense of a committee.

| **utvalgN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 50/71 | .70 |
| Sense distribution: | | | |
| **Sense1:** | {*range selection variety*} | 22/50 | .44 |
| **Sense3:** | {*committee*} | 23/50 | .46 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense2:** | {*assortment*} | 2/50 | .04 |
| **Sense4:** | {*county**} | 1/50 | .02 |
| **Sense5:** | {*gallery*} | 1/50 | .02 |
| **Sense6:** | {*sample*} | 1/50 | .02 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.4: *utvalg*N automatically sense-tagged material and sense distribution.

The target noun *utvalg*N is thus interesting because on the one hand, the sense distinction between the Mirrors sense 1 and 2 appears quite clear-cut (although the senses are related), and also the senses are quite evenly distributed in the training material.

### *rot*N

*rot*N was deemed suitable because its sense divisions (considering the sufficiently frequent ones) fit quite well with those of Bokmålsordboka, which defines a homonymous distinction between 'root, origin' and 'mess, disorder'. Both meanings may have a concrete and a metaphoric interpretation. Their semantic unrelatedness makes them distinct lexemes, which also surfaces through the fact that they

have the same lexical lookup form but belong to different inflectional paradigms. The 'mess' lexeme is a neuter noun and the 'root' lexeme is a masculinum noun, which means in Norwegian that they belong to different inflectional patterns. Since we only consider the lemma form, however, this morphological difference is not taken into account.

***rot*N senses according to Bokmålsordboka**

I. root (five subsenses comprising for instance the physical root as well as the metaphoric notion of 'origin')

II. mess, unorder (physically (*rommet var fullt av rot* 'the room was full of mess') or in the sense of an unorderly or confusing situation)

| **rotN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 62/92 | .67 |
| Sense distribution: | | | |
| **Sense2:** | {*origin root*} | 43/62 | .69 |
| **Sense7:** | {*mess*} | 10/62 | .16 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense1:** | {*clutter confusion*} | 2/62 | .03 |
| **Sense3:** | {*Mrs**} | 1/62 | .02 |
| **Sense4:** | {*base*} | 3/62 | .05 |
| **Sense5:** | {*bedroom**} | 1/62 | .02 |
| **Sense6:** | {*disorder*} | 1/62 | .02 |
| **Sense8:** | {*problem*} | 1/62 | .02 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.5: Mirrors entry. *rot*N automatically sense-tagged material and sense distribution.

### *plan*N

According to Bokmålsordboka, this noun has two main senses:

***plan*N senses according to Bokmålsordboka**

I. project, schedule, plan

II. level

This sense distinction matches with the two sufficiently frequent senses in the Mirrors method (Table (8.6) (p. 175)), in which sense 1 denotes the concept of 'programme, project, schedule, scheme' (*etter planen* 'according to schedule') and sense 2 denotes 'level' (*på det regionale plan* 'on the regional level').

| **planN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 118/170 | .69 |
| Sense distribution: | | | |
| **Sense1:** | {*programme project schedule scheme*} | 94/118 | .80 |
| **Sense2:** | {*level plane*} | 19/118 | .16 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense3:** | {*design**} | 1/118 | .01 |
| **Sense4:** | {*fanfare**} | 1/118 | .01 |
| **Sense5:** | {*pace*} | 1/118 | .01 |
| **Sense6:** | {*plan*} | 0/118 | .00 |
| **Sense7:** | {*planning*} | 1/118 | .01 |
| **Sense8:** | {*stand**} | 1/118 | .01 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.6: Mirrors entry: *plan*N automatically sense-tagged material and sense distribution.

### *valg*N

Consulting Bokmålsordboka, the Norwegian noun *valg*N only has one main sense which is divided in two sub-senses. Both senses pertain to choosing or selecting, but whereas the first sub-sense concerns the act of making a choice or voicing a preference (e.g. *etter eget valg* 'of your own choice'), the second denotes specifically the act of electing somebody for a position (*election, ballot*).

***valg*N senses based on Bokmålsordboka**

I. 1 choice

2 election

One may thus argue that there is no homonymy in the case of *valg*. Nonetheless it was selected as a target noun for WSD classification because its sense distinction in the Mirrors word base is perfectly consistent with the decision of a common Norwegian dictionary; the lemma is relatively frequent in the corpus, and the sense distribution is quite satisfactory in terms of the number of instances per sense.

| **valgN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 119/150 | .79 |
| Sense distribution: | | | |
| **Sense1:** | {*choice*} | 44/119 | .37 |
| **Sense2:** | {*election*} | 74/119 | .62 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense3:** | {*gain**} | 1/119 | .01 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.7: Mirrors entry for *valg*N: automatically sense-tagged material and sense distribution.

As the table shows, the Mirrors generated three senses of *valg*N, of which only two were sufficiently frequent in the corpus according to our minimum frequency threshold (cf. p. 166). The sufficiently frequent senses are represented by the sense partitions {choice} and {election}, respectively, which intuitively covers the same sense distinction as we saw in Bokmålsordboka. The latter of these senses is more dominant in the training corpus than the other (74 vs. 44 instances).

### *slag*N

The Norwegian noun *slag* was selected as a target noun for WSD classification because its sense distinctions accord quite well with the common dictionary Bokmålsordboka, the lemma is relatively well represented in the corpus, and it is interesting to include since it is more than two-ways ambiguous.

According to Bokmålsordboka, *slag*N may be divided into three main senses based on their etymological origin.

| ***slag*N senses in Bokmålsordboka** |
|---|
| I. hitting (in a wide sense) |
| II. type, kind |
| III. a small, additional outdoor building to keep for instance doves or wasps |

For the sake of a quick overview, all subsenses listed in Bokmålsordboka are not listed in the framed overview above. The first main sense pertains to 'hitting' in a very wide sense, comprising for instance the subsenses of a battle, giving or receiving a blow in a physical as well as a metaphoric sense, or having a stroke (in the sense of an illness). The second sense refers to a 'kind, type' (*kaker av mange slag* 'cakes of many kinds'). The last sense listed in the dictionary is a quite specific sense which refers to a small, additional outdoor building to keep for instance doves or wasps; this sense was never observed in the ENPC.

In the Mirrors method no less than twelve sense partitions were generated for *slag*N (Table (8.8) (p. 177)), but the majority of these are one-instance senses, and a manual verification revealed that many of these reflect errors from automatic word alignment. Four of the senses satisfy the minimum frequency threshold of 10 sense-tagged corpus instances; these are listed first. The remaining senses are discarded from further consideration.

The four sufficiently frequent senses match well with the *a priori* sense inventory of Bokmålsordboka: The Mirrors sense 1 nicely teams up with Bokmålsordboka's sense 2 ('kind, sort, type'), whereas the Mirrors senses 2, 3 and 12 encapsulate various subsenses of sense 1 in Bokmålsordboka ('battle, blow, stroke'). Since Bokmålsordboka predicts that *slag*1 (the 'kind' sense) is the only

| **slagN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 148/206 | .72 |
| Sense distribution: | | | |
| **Sense1:** | {*age* ease* fish* kind living* manner nature sort style type*} | 86/148 | .58 |
| **Sense2:** | {*battle*} | 22/148 | .15 |
| **Sense3:** | {*blow*} | 19/148 | .13 |
| **Sense12:** | {*stroke*} | 12/148 | .08 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense4:** | {*cash**} | 1/148 | .01 |
| **Sense5:** | {*collar*} | 1/148 | .01 |
| **Sense6:** | {*litany**} | 1/148 | .01 |
| **Sense7:** | {*punch*} | 2/148 | .01 |
| **Sense8:** | {*servant**} | 1/148 | .01 |
| **Sense9:** | {*silence**} | 1/148 | .01 |
| **Sense10:** | {*size**} | 1/148 | .01 |
| **Sense11:** | {*strike*} | 1/148 | .01 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.8: Mirrors entry: *slag*N automatically sense-tagged material and sense distribution.

sense which is clearly separate from the other Mirrors senses, the Mirrors might have grouped senses 2, 3 and 12 together according to etymological origin. As will be discussed when considering the training and test material (Section (8.3)), this was in part the experience when manually verifying the automatically sense-tagged instances and manually sense-tagging the test instances: senses 1 and 2 were easy to separate from each other and from the last two, but senses 3 and 12 proved to overlap in meaning.

### *tak*N

Bokmålsordboka lists two main senses of this noun, i.e. it covers two homonyms.

| ***tak*N senses according to Bokmålsordboka** |
|---|
| I. roof, ceiling, shield, cover |
| II. grip, hold (also in the metaphoric sense of having the power over someone or something) |

The Mirrors generated 12 senses altogether, of which three were sufficiently frequent: sense 2 ('grasp, hold'), and the two similar meanings of sense 4 ('ceiling') and sense 7 ('roof'). This TW was included under doubt, since—intuitively—the two most frequent Mirrors senses should perhaps actually have been grouped together. In other words, since it is quite likely that a WSD should confuse these two senses, it is likely that the precision of the classifier may appear to be quite low for this target word. Hopefully, however, we may still analyse the results

in a useful way by considering the so-called confusion matrices (checking which senses were confused with each other).

| **takN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 152/380 | .40 |
| Sense distribution: | | | |
| **Sense2:** | {*grasp hold*} | 16/152 | .11 |
| **Sense4:** | {*ceiling*} | 52/152 | .34 |
| **Sense7:** | {*roof*} | 74/152 | .49 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense1:** | {*Lake* lake**} | 2/152 | .01 |
| **Sense3:** | {*back**} | 1/152 | .01 |
| **Sense5:** | {*mass**} | 1/152 | .01 |
| **Sense6:** | {*pot**} | 1/152 | .01 |
| **Sense8:** | {*sky**} | 1/152 | .01 |
| **Sense9:** | {*stroke*} | 1/152 | .01 |
| **Sense10:** | {*tail**} | 1/152 | .01 |
| **Sense11:** | {*top*} | 1/152 | .01 |
| **Sense12:** | {*truck**} | 1/152 | .01 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.9: Mirrors entry: *tak*N automatically sense-tagged material and sense distribution.

### *stemme*N

According to Bokmålsordboka, this noun has one main sense which is divided into three subsenses:

| ***stemme*N senses according to Bokmålsordboka** |
|---|
| I. 1 voice |
| 2 a particular voice in music; e.g. the soprano voice |
| 3 a vote in elections |

The Mirrors generated three sense partitions for *stemme*N, sense 2 being a one-instance sense that resulted from an erroneous word alignment between *stemme* and *policeman* (the context was: *Politimannens stemme ble...* 'the policeman's voice became...'). The training material has a relatively high total number of instances, although with a very skewed sense distribution: the most frequent sense (sense 1, 'voice') accounts for $94\%$ of the material.

### *liv*N

This noun was included under doubt since on the one hand, its ambiguity in the Mirrors intuitively appears quite clear, but on the other hand the sense distribution is extremely skewed.

| **stemmeN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 384/579 | .66 |
| Sense distribution: | | | |
| **Sense1:** | {*brother* father* mother* tone tune voice*} | 360/384 | .94 |
| **Sense3:** | {*vote*} | 23/384 | .06 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense2:** | {*policeman**} | 1/384 | .00 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.10: Mirrors entry: *stemme*N automatically sense-tagged material and sense distribution.

Following Bokmålsordboka, the noun *liv*N is not ambiguous in terms of homonymy, although some of its related senses are more clearly distinct than others. Bokmålsordboka lists no less than 14 subsenses. We will not list all of them here instead they will be summed up as follows:

senses 1–8 essentially comprise a notion of life as an abstract entity of some kind: 'existence', 'a lifetime', 'way of living' (*a peaceful life*), 'area of human activity' (*sex life, family life*), 'amusement, fun' (*to live life*), 'energy' (*full of life*), 'some living creature' (*plant life, animal life*), and 'a next life after death' (*eternal life*)

The dictionary senses 9–14 have a bodily reference in various ways: 'bodily life' (as opposed to spiritual life, as in *with his life and soul*), 'the lower part of the torso', 'belly', 'waist', 'womb', and finally *liv* as a the torso part of a piece of clothing (*e.g. a dress with a red upper part*).

The Mirrors generated 9 senses (Table (8.11) (p. 180)), but only two of them exceed our frequency threshold. The most frequent sense (disregarding erroneous word alignments, marked by an asterisk in the table) neatly matches the 'being' senses 1–8 in the Bokmålsordboka dictionary, whereas the least frequent sense points to the 'lower part of torso' sense of 'waist',

## 8.4.2 Target adjectives

### *full*AJ

Consulting Bokmålsordboka, four sub-senses are enumerated:

| ***full*AJ senses according to Bokmålsordboka** | |
|---|---|
| I. 1 | with maximal content (*et fullt glass* 'a full glass') |
| 2 | complete, unlimited |
| 3 | completely, in every respect |
| 4 | drunk (alcohol) |

| **livN** | | | |
|---|---|---|---|
| Total sense-tagged: | | 1048/1334 | .79 |
| Sense distribution: | | | |
| **Sense1:** | {*Life being land* life lifestyle lifetime living middle* mind* movement* spirit time* will**} | 1023/1048 | .98 |
| **Sense8:** | {*waist*} | 17/1048 | .02 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense2:** | {*Aid**} | 1/1048 | .00 |
| **Sense3:** | {*Olav**} | 1/1048 | .00 |
| **Sense4:** | {*bodice*} | 2/1048 | .00 |
| **Sense5:** | {*hip*} | 1/1048 | .00 |
| **Sense6:** | {*limit*} | 1/1048 | .00 |
| **Sense7:** | {*pain**} | 1/1048 | .00 |
| **Sense9:** | {*womb*} | 1/1048 | .00 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.11: Mirrors entry: *liv*N automatically sense-tagged material and sense distribution.

The adjective *full*AJ has been divided in six main senses according to the Mirrors method, of which only two exceed our frequency threshold (Table (8.12) (p. 180)). The Mirrors sense 1 denotes the same notion as expressed in Bokmålsordboka's senses 1, 2 and 3, viz. 'full, complete'. The Mirrors sense 2 encapsulates Bokmålsordboka's sense 4, that of 'being drunk'.

| **fullAJ** | | | |
|---|---|---|---|
| Total sense-tagged: | | 308/632 | .49 |
| Sense distribution: | | | |
| **Sense1:** | {*able busy complete entire full little* loud much powerful*} | 280/308 | .91 |
| **Sense2:** | {*drunk drunken*} | 25/308 | .08 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense4:** | {*dull*} | 1/308 | .00 |
| **Sense5:** | {*full-time*} | 1/308 | .00 |
| **Sense6:** | {*respectful*} | 1/308 | .00 |
| **Sense3:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.12: Mirrors entry: *full*AJ automatically sense-tagged material and sense distribution.

### *gal*AJ

The adjective *gal*AJ has been divided in two main senses according to the Mirrors method, viz. 'crazy' and 'incorrect' (Table (8.13) (p. 181)).

Consulting Bokmålsordboka, five sub-senses are enumerated:

| ***gal*AJ senses according to Bokmålsordboka** | |
|---|---|
| I. | 1 insane, crazy |
| | 2 strongly interested in/in love with |
| | 3 incorrect, spurious |
| | 4 unfortunate, wrong |
| | 5 illegal, wrong |

Considering the senses listed in Bokmålsordboka, senses 1–2 encompass a notion of wildness or craziness: 'mad, crazy' in sub-sense 1 and 'wildly, uncontrollably' (*madly in love*) in 2. This corresponds nicely to the Mirrors sense *gal*1 ('crazy, insane, mad'). Senses 3–5 refer in various ways to 'something not right'. Senses 3 and 5 relate to 'incorrectness' in the objective sense and in a moral sense, respectively (*the wrong(incorrect) direction* and *stealing something is wrong*). These two senses together thus constitute a clear counterpart to the Mirrors sense *gal*2 ('false, incorrect, wrong'). Sense 4, denoting 'bad consequences' (*a bad situation*) resembles the Mirrors sense *gal*3 ('bad'), which was unfortunately discarded on statistical grounds.

| **galAJ** | | | |
|---|---|---|---|
| Total sense-tagged: | | 124/178 | .70 |
| Sense distribution: | | | |
| **Sense1:** | {*crazy insane mad*} | 36/124 | .29 |
| **Sense2:** | {*false incorrect wrong*} | 82/124 | .66 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense3:** | {*bad*} | 6/124 | .05 |

∗Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.13: Mirrors entry: *gal*AJ automatically sense-tagged material and sense distribution.

This adjective has a fairly good frequency, given the ENPC as our source, although the sense distribution is relatively skewed.

### *frisk*AJ

According to Bokmålsordboka, *frisk*AJ is listed with six subsenses:

| ***frisk*AJ senses according to Bokmålsordboka** | | |
|---|---|---|
| I. | 1 | non-faded, pure (*friske egg, friske blomster* 'fresh eggs, fresh flowers') |
| | 2 | new, renewed (*friskt blod* 'fresh blood (new people') |
| | 3 | new, recent (*ferske spor* 'recent leads') |
| | 4 | strong, lively, healthy (*frisk bris* 'strong wind', *friske ungdommer* 'strong/healthy youth') |
| | 5 | refreshing (*en frisk smak* 'a refreshing taste') |

Considering the Mirrors entries, four entries were generated of which two are above the frequency threshold: sense 1 alludes to the notion of 'new' (as in the Bokmålsordboka senses 1, 2 and 3). The Mirrors sense 2 is ignored, being a BAG-OF-SINGLETONS partition (Chapter (4.4.3)), and sense 3 is not sufficiently frequent. Sense 4 ('healthy') is close to the Bokmålsordboka sense 4.

| **friskAJ** | | | |
|---|---|---|---|
| Total sense-tagged: | | 66/119 | .55 |
| Sense distribution: | | | |
| **Sense1:** | {*bright fresh good green new sweet*} | 47/66 | .71 |
| **Sense4:** | {*healthy*} | 18/66 | .27 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense3:** | {*brisk*} | 1/66 | .02 |
| **Sense2:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified).

Table 8.14: Mirrors entry: *frisk*AJ automatically sense-tagged material and sense distribution.

### 8.4.3 Target verbs

#### *trykke*V

The verb *trykke*V has two etymologically unrelated senses according to Bokmålsordboka.

| ***trykke*V senses according to Bokmålsordboka** | | |
|---|---|---|
| I. | 1 | print (*trykke en bok* 'print a book') |
| | 2 | print a pattern (*printe mønster på en stol* 'print a pattern on a chair') |
| II. | 1 | push, squeeze (*trykke på en knapp* 'push on a button') |

The former sense pertains to 'printing' (text or patterns on paper or other material), whereas the latter denotes the activity of 'squeezing together or pushing'.

This sense distinction teams up perfectly with the sense inventory derived by the Mirrors method, as can be seen in Table (8.15) (p. 183).

| **trykkeV** | | | |
|---|---|---|---|
| Total sense-tagged: | | 41/68 | .60 |
| Sense distribution: | | | |
| **Sense1:** | {*capture* carry have*[?] *press receive shake*[?] *squeeze take*} | 31/41 | .76 |
| **Sense2:** | {*print*} | 10/41 | .24 |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified). Instances where no correspondence at all could be found manually are marked by a raised question mark ([?]).

Table 8.15: Mirrors entry: *trykke*V automatically sense-tagged material and sense distribution.

Frequency clearly poses a problem, in total the two senses are only represented by 41 instances; additionally the sense distribution is rather skewed. The verb is still included since the Mirrors sense distinctions are optimal and since we may use the verb for testing if there is any improvement in learning when abstracting from (sparse) context words to semantic-features.

### *utsette*V

The verb *utsette*V is not homonymous in terms of etymology, but intuitively the Mirrors senses that satisfy our minimum frequency threshold still appear to be clearly separate. According to Bokmålsordboka, *utsette*V has four related senses:

***utsette*V senses according to Bokmålsordboka**

I. 1 postpone (*utsette møtet* 'postpone the meeting')

2 criticise (*ha litt å utsette på naboene* 'have things to criticise the neighbours for')

3 expose (*utsette seg for fare* 'expose oneself to danger')

4 (music:) arrange (*melodien ble utsatt for orkester* 'the melody was arranged for an orchestra')

The concept of 'postponing' finds its clear counterpart in the Mirrors sense *utsette*3 (Table (8.16) (p. 184)). Sense 2 of the verb in Bokmålsordboka points to the notion of 'criticising', and is only realised in the infinitive when the verb is followed by the preposition *på* 'on' (phrasal verb). A quick search in the ENPC revealed that this phrase only occurs once in the entire corpus, and in the corresponding sentence the verb phrase has not been given a direct translation: *Stille, ordentlige mennesker som ingen hadde noe å utsette på. (EG2)* 'Quiet, ordinary, respectable men, both of them.'. Bokmålsordboka's sense 3 refers to 'exposing (smd. to smth.)', which nicely corresponds to the Mirrors sense *utsette*1. The

fourth dictionary sense is a rather specialised, and hence marginal, musical sense of 'arranging (a melody)'. This sense was not attested in the ENPC.

This verb actually displays an error, not related to the word alignment, namely that the Mirrors *utsette*1 contains the English translation *suspend*, which should more appropriately be grouped with the postponing sense *utsette*3. It has not been clarified why this correspondent was grouped with sense 1.

| **utsetteV** | | | |
|---|---|---|---|
| Total sense-tagged: | | 45/115 | .39 |
| Sense distribution: | | | |
| **Sense1:** | {*endure* expose have*$^{?}$ *keep* leave* put* strengthen* stress* suffer* suspend tolerate**} | 23/45 | .51 |
| **Sense3:** | {*delay postpone*} | 18/45 | .40 |
| Senses w/ frequency $\leq 10$: | | | |
| **Sense4:** | {*discredit**} | 1/45 | .02 |
| **Sense5:** | {*erode**} | 1/45 | .02 |
| **Sense6:** | {*subject*} | 2/45 | .04 |
| **Sense2:** | BAG-OF-SINGLETONS | — | — |

*Starred lemmas are false translational correspondents from aut. word alignment (manually verified). Lemmas marked by $^{?}$ indicate that it was not ascertained where the target word found to correspond to this lemma

Table 8.16: Mirrors entry: *utsette*V automatically sense-tagged material and sense distribution.

Neither of the senses that the Mirrors did identify are very frequent (only $45$ training instances altogether), but although not frequent, the two senses have a relatively even sense distribution. The verb was therefore found to constitute an interesting target verb for WSD.

## 8.5 A manual inspection of the data sets

### General

This section evolves around the work of manually sense-tagging target word instances that were not automatically sense-tagged (using a shorter term, we will refer to these as the 'untagged' instances). This work gave rise to observations about the quality of the automatically sense-tagged as well as the untagged data sets in general.

Any manual sense-tagging presupposes that the annotator has acquainted herself with the given sense inventory, be it a WordNet sense inventory or a Mirrors sense inventory. As presented in the first subsection below, the basis for interpreting the Mirrors sense inventory was to consider the concrete examples from the automatically sense-tagged material. The same section presents some observations about the quality of the automatically sense-tagged data sets.

In the subsequent subsection the principles for manual sense-tagging are presented and discussed: as we will see, it was in the set of untagged data that pre-processing errors and other unwanted target word instances were found, so some of the instances had to be discarded. The amounts of and the manually estimated quality of the automatically sense-tagged data are summed up in Figure (8.17) (p. 191); the amounts of manually sense-tagged data sets are summed up in Figure (8.18) (p. 192). The joint data sets are summed up in Figure (8.19) (p. 193).

**The automatically sense-tagged material**

An important property of translational correspondences is that they are often surprising, and yet remarkably plausible when considering the situated context. This was often the experience when the plausibility of each translational correspondent of a target word was verified manually by the author in Section (8.4).

Since the use of corpus data sometimes generates unanticipated and yet plausible information, it was decided to interpret the set of sufficiently frequent senses as a 'given', even in the case that the sense divisions do not appear to be optimal at first glance. That is, it was decided to at least consider the possibility that the difference (or the lack of difference) between Mirrors senses are motivated when considering the contexts that gave rise to various translational choices. Therefore, before sense-tagging the untagged instances, it was chosen to first consider the automatically sense-tagged instances in order to get an idea of how the Mirrors senses are used.

Based on a scrutiny of the automatically sense-tagged instances, most of the target words (*fyr*N, *utvalg*N, *rot*N, *plan*N, *valg*N, *stemme*N, *liv*N, *gal*AJ, *full*AJ, *trykke*V and *utsette*AJ) were quite unproblematic (judging the sufficiently frequent senses only). Their sense divisions were quite easy to grasp and accordingly the subsequent manual sense-tagging of untagged instances (to be discussed in the next subsection) becomes quite straightforward.

Still, there were a few cases of erroneous automatic sense-tags. These errors are recorded in Table (8.17), which sums up the quality of the automatically sense-tagged material, based on a manual verification. As the table shows, such cases were quite rare: one error was observed in the *plan*N material, one in *tak*N, one in *trykke*V, three errors in the case of *full*AJ and five errors in the case of *utsette*V.

Such errors resulted from word-alignment errors and not from the Mirrors method in itself. As an example, an instance of *utsette*V in the sense of 'postpone' was (erroneously) word-aligned with *strengthen*. Example (7) shows the full sentence and the corresponding English sentence from the ENPC (in fact, the Norwegian sentence has been found to correspond to two English sentences, which is probably what caused the word-alignment error). Intuitively, this instance should

have been word-aligned, not with *strengthen* (the last but three words in the second sentence) but with *postpone* (the 18th word in the first sentence). The recorded *strengthen* alignment happened to be grouped as a member of the sense partition representing the concept 'to expose' (i.e. sense 1 in Figure (8.16) (p. 184)), meaning that the instance did not receive the intuitively desirable sense-tag.

(7) Magda hadde naturligvis protestert så iherdig at han nesten hadde **utsatt** slankingen til mandag, men så hadde han sett frøken Borgs medlidende smil for seg og vært standhaftig. (EG2)
He had started by renouncing the previous evening's steak, despite his decision a few hours earlier to postpone his change of diet one more day. True to form, Magda had protested and urged him to wait till Monday, but the memory of Miss Borg's condescending smile had strengthened his resolve.

### Manual sense-tagging

As for the set of instances that were not sense-tagged automatically—what we with a shorter (and less precise) term may call the *untagged* instances—they are untagged in virtue of not being word-aligned. We may roughly categorise the untagged instances as belonging to three categories, where those belonging to the former two were not included in the final data sets:

1. lemmatisation errors (excluded from the final data set)
2. idiomatic (fixed) expressions (excluded from the final data set)
3. instances with a non-literal translational correspondence (in the final data set)

A bit simplified, the two first categories represent cases of 'atypical' analyses or uses of the target word, whereas the last category result from atypical choices on the part of the *translator*—most of the untagged instances, in fact, fell in to the third category, exemplifying the well-known phenomenon that translators do not always choose a literal rendering of the source sentence; and in that case a successful word alignment (as well as a successful automated sense-tagging) is barred.

An interesting question before considering the untagged instances was whether they reveal word senses that were not at all covered by the Mirrors method. The short answer is 'no': Specifically, the answer is 'no' if one accepts a general distinction between general senses of a word as opposed to idiomatic expressions. Such a distinction is justified because idiomatic expressions count as such precisely because it is the (multiword) *expression*, and not the individual

words in the expression, that has a meaning (a so-called non-compositional meaning). Since idiomatic expressions usually do not have a word-by word translation, they are, unsurprisingly, hard to find in the translation-based (and individual word-oriented) Mirrors method. It *did* occur that the dictionary Bokmålsordboka sometimes listed senses that were not covered at all in the Mirrors method (cf. the discussion of *slag*N in Section (8.4.1) (p. 176)). But when a sense predicted by Bokmålsordboka was not found in the Mirrors sense partitions, it was also not found among the untagged instances.

In order to annotate the untagged instances manually, they were first printed to a file showing their immediate sentence context (in the few cases where more than the sentence context was needed, the author simply looked up the instance in the corpus where the full document context can be accessed). An instance was tagged or discarded as follows:

- Instances that match a sufficiently frequent sense (according to the training material counts) are tagged with their appropriate sense-tag.
- Instances are discarded under the following circumstances:
  - (i) Instances with an incorrect analysis (not being the target word) (cf. Example (8)).
  - (ii) Instances where the target word is clearly part of fixed expressions and idioms (cf. Example (9)).
  - (iii) Instances that match an insufficiently frequent sense (according to the training material counts) (cf. Example (10)).
- If more than one sufficiently frequent sense-tag could be appropriate for a test instance, the most frequent of the senses (according to the frequencies from the training material) is chosen[4] (Example (11)).

Some of the untagged instances exemplified preprocessing errors, that is, lemmatisation errors: If the lemmatiser erroneously tagged a verb as a noun, and the verb corresponds to a verb in the corresponding sentence, the word-aligner cannot succeed since it only word-aligns two words that belong to the same part of speech. Example (8) shows a lemmatisation error, where an instance lemmatised as *stemme*N really was an instance of the verb *stemme*V.

Example (9) illustrates a fixed expression, *(være) i slag* '(be) on the rise', which is discarded from the set of test instances since intuitively, the individual words of the expression cannot be said to have an independent meaning. As this

[4]This does not mean that the most frequent sense of all target word senses is chosen, but the most frequent of those senses that are not sufficiently distinct.

sentence example also illustrates a case of truly non-literal rendering of the original sentence (the Norwegian sentence) on the part of the translator, the Norwegian sentence is given a word-by-word gloss, then the corresponding English sentence (as found in the corpus) is given.

Example (10) illustrates an instance of *lag*N in the untagged data set that is lost because none of the sufficiently frequent senses are relevant. The two sufficiently frequent senses are *lag*5 ('layer') and *lag*8 ('team'), but none of these plausibly cover the meaning of 'party, company', which was covered by the insufficiently frequent Mirrors sense of *lag*1.

(8) **Stemmer** nok det, ja, svarte mannen. (MM1T)
"Sure as eggs is eggs", said the man.

(9) Hun kan det der når hun er i slag. (JM1)
She can all that when she is on the rise
She likes to analyse.

(10) Om aftenen inviterer de lokale yachtklubber til godt **lag** i sine klubbhus. (KT1)
In the evening the local yacht clubs invite everyone to parties in their clubhouses.

Three of the target words revealed questionable sense divisions that posed problems when attempting to sense-tag test instances with these sense inventories. Each of these will be discussed in what follows.

### Questionable sense divisions (i): *slag*N

The lemma *slag*N is interesting because more than two senses exceeded the frequency threshold. A scrutiny of the automatically sense-tagged training instances revealed that no errors were identified, but the affinity between its Mirrors senses 3 ('blow') and 12 ('stroke') was striking. Sense 3 ('blow') covers both the physical 'strike, blow' and, metaphorically, a defeat. Sense 12, based on the automatically sense-tagged material, mostly pointed to the sense of a heart attack (*få slag* 'have a heart attack'), a sense which (at least intuitively) is easily separated from those in sense 3. But also, the automatic sense-tagging also assigned sense 12 to two cases of metaphor (*ikke/aldri gjøre et slag* 'not/never do a stroke' and *med ett slag* 'at a stroke') and two instances that would intuitively be just as close to the sense *slag*3: the one pertains to heart beats (*[hjertet] slo med forte, syke slag* '[his heart] beat with quick, feeble strokes') and the other has to do with the beat of wings (*nedoverslaget er det **slaget** som gir mest kraft* 'the power **stroke** is the downward beat of the wings').

This makes it hard to conceive of a fully convincing difference between senses 3 and 12 of *slag*N; which, in turn, made it difficult to choose between the senses

when sense-tagging instances of *slag*N. Given for instance a context about heart beats (*hoppet over flere slag* 'skipped several beats'), it could intuitively count as *slag*3 ('blow') but the automatically sense-tagged material predicted *slag*12 ('stroke') in these cases. In line with the sense-tagging principles stated above, we then choose the sense 3 and not sense 12 because the former is more frequently seen in the training material. Untagged instances that clearly pertained to a heart attack, though, were tagged as *slag*12 ('stroke').

**Questionable sense divisions (ii): *tak*N**

As for *tak*N, the difference between *tak*4 ('ceiling') and *tak*7 ('roof') is clearly marginal for Norwegian, in that the former would indicate the inside of a roof and the latter the outside, a distinction which is not lexicalised in Norwegian for *tak*N. This was thus the perhaps best example of the need to decide to just take the Mirrors sense inventory as a given. Accordingly, in cases where the context describes, say, birds flying over the roof, then it is easy to choose sense *tak*7. Other contexts exemplify cases of under-specification, where it is not clear (or not relevant) if the inside or the outside is meant, for instance in the expression *tak over hodet* 'a roof over one's head'; one example from the ENPC being given in Example (11). In the example, the Norwegian sentence was the translation of the English original, and the Norwegian expression *tak over hodet* 'a roof over one's head' corresponds to *shelter* in the given English original sentence.

During the manual sense-tagging this was solved in the following way: In cases of under-specification, the most frequent of the polysemous senses is chosen as default, in our example *tak*7. Note that we then choose the most frequent sense among those senses about which there is doubt. In our example where *tak*4 and *tak*7 are clearly polysemous, the frequency of *tak*2 is not considered since this sense is easily separated from the two others.

(11) (. . . ) og finner ikke **tak** over hodet, mat og venner andre steder enn der. (LTLT1T)
(. . . ) and their only source of shelter, food and companions

**Questionable sense divisions (iii): *frisk*AJ**

For the adjective *frisk*N, three questionable sense-tags were found in the automatically sense-tagged material that were not technically spurious, but they illustrate that the sense divisions are not perfect. Example (12) shows an instance where *frisk*N occurs in the sense of not 'new' but 'healthy, functioning'. Intuitively, therefore, I would expect the Mirrors sense 2 ('healthy') here. As it turned out, however, the TW instance was (correctly) word-aligned with English *good*, which

is a member of the sense partition of *frisk*1 ('fresh, good, new'). Technically, the instance is thus correctly sense-tagged (and we do not register it as an error in the table counts), but we make a note that the sense distinction is not unproblematic.

(12) Han bløddе fra den **friske** hånden, men (...) (HW1)
He bled from the **good** hand, but (...)
His good hand was bleeding but (...)

**Summing up the automatically and manually sense-tagged test sets**

Beginning with the automatically sense-tagged material, Table (8.17) (p. 191) below sums up the results from the manual verification of this material.

Providing counts per sense, we borrow the nomenclature used in confusion matrices of classification results and use the terms 'true positives' (TP) and 'false positives' (FP) (Daelemans et al., 2007, p. 31). The 'true positives' (TP) is the number of instances correctly classified as a particular class $c$; 'false positives' (FP) is the number of instances that were incorrectly tagged as $c$. In other words, TP + FP together count how many of the instances were tagged as a sense $c$; TP alone counts how many of these were correct.

So in Table (8.17) (p. 191) we use these terms as follows: If all instances of a sense were deemed to be correct (i.e. TP=TP + FP) only the total count is given in the third column (TP + FP). If the TP count is lower than the total, the relative frequency is given (TP / TP + FP). The last row returns the Overall Accuracy, OA, based on the counts in the third row. The OA is given by the numerical relative frequency (#) (the number of correct sense-tagged seen relative to the total number of sense-tagged instances) and by its corresponding percentage (%).

The manually sense-tagged material is summed up in Table (8.18) (p. 192), showing the sense distribution, given the sufficiently frequent senses, in the manually sense-tagged material. The third column shows the number of sense-tagged instances per sense. The fourth and fifth column show the *a priori* sense probability of each sense (i.e. P(s)) in the automatically and manually sense-tagged material, respectively. Thus the interested reader may easily compare the sense distributions in the automatically and manually sense-tagged sets. The most frequent sense (MFS) is marked in bold for both data sets per word.

As Table (8.18) shows, most of the target words have the same MFS in both data sets; the exceptions being *valg*N, *slag*N and *tak*N (this deviation is discussed in the following section that sums up and discusses the presented lexical sample data set). Note that the P(s) was calculated as a relative frequency counting the frequency of each sense given the total number of TW instances tagged *with a sufficiently frequent sense* (cf. also Equation (6.3), p. 126). Therefore, the P(s) for the training material in Table (8.18) will differ slightly from the corresponding relative frequencies in the Mirrors entries for each target word in Section (8.4) (the

Table 8.17: 15 target words, **automatically sense-tagged data set**. Manually verified accuracy of the automatic sense-tagger (training material)

| | Lemma | Senses | TP+FP | OA: # | % |
|---|---|---|---|---|---|
| Nouns | *lag*N | *lag*5 ('layer') | 23 | 36/36 | 100% |
| | | *lag*8 ('team') | 13 | | |
| | *fyr*N | *fyr*1 ('guy') | 36 | 48/48 | 100% |
| | | *fyr*2 ('fire') | 12 | | |
| | *utvalg*N | *utvalg*1 ('selection') | 22 | 45/45 | 100% |
| | | *utvalg*3 ('committee') | 23 | | |
| | *rot*N | *rot*2 ('root') | 43 | 53/53 | 100% |
| | | *rot*7 ('mess') | 10 | | |
| | *plan*N | *plan*1 ('scheme') | 94 | 112/113 | 99.1% |
| | | *plan*2 ('level') | 18/19 | | |
| | *valg*N | *valg*1 ('choice') | 44 | 118/118 | 100% |
| | | *valg*2 ('election') | 74 | | |
| | *slag*N | *slag*1 ('kind, type') | 86 | 139/139 | 100% |
| | | *slag*2 ('battle') | 22 | | |
| | | *slag*3 ('blow') | 19 | | |
| | | *slag*12 ('stroke') | 12 | | |
| | *tak*N | *tak*2 ('grasp') | 16 | 140/142 | 98.6% |
| | | *tak*4 ('ceiling') | 52 | | |
| | | *tak*7 ('roof') | 72/74 | | |
| | *stemme*N | *stemme*1 ('voice') | 360 | 383/383 | 100% |
| | | *stemme*3 ('vote') | 23 | | |
| | *liv*N | *liv*1 ('life') | 1023 | 1040/1040 | 100% |
| | | *liv*12 ('waist') | 17 | | |
| Adjectives | *full*AJ | *full*1 ('complete') | 277/280 | 302/305 | 99.01% |
| | | *full*2 ('drunk') | 25 | | |
| | *gal*AJ | *gal*1 ('crazy') | 36 | 118 | 100% |
| | | *gal*2 ('incorrect') | 82 | | |
| | *frisk*AJ | *frisk*1 ('fresh') | 47 | 65/65 | 100% |
| | | *frisk*4 ('healthy') | 18 | | |
| Verbs | *trykke*V | *trykke*1 ('squeeze') | 30/31 | 40/41 | 97.56% |
| | | *trykke*2 ('print') | 10 | | |
| | *utsette*V | *utsette*1 ('expose') | 18/23 | 36/41 | 87.80% |
| | | *utsette*3 ('postpone') | 18 | | |

latter are based on the total number of sense-tagged instances, that is, including also the insufficiently frequent senses).

Table 8.18: 15 target words, **manually sense-tagged data set**. The *a priori* sense probabilities (P(s) of each sense in the automatically vs. manually sense-tagged (AST vs. MST) data sets are both listed for comparison. The most frequent sense in both data sets per target word is marked in bold.

| | Lemma | Senses | Tags | Total instances | P(s) AST | P(s) MST |
|---|---|---|---|---|---|---|
| Nouns | *lag*N | *lag*5 ('layer') | 14 | 18 | **.64** | **.78** |
| | | *lag*8 ('team') | 4 | | .36 | .22 |
| | *fyr*N | *fyr*1 ('guy') | 26 | 34 | **.75** | **.76** |
| | | *fyr*2 ('fire') | 8 | | .25 | .24 |
| | *utvalg*N | *utvalg*1 ('selection') | 8 | 21 | .49 | .38 |
| | | *utvalg*3 ('committee') | 13 | | **.51** | **.62** |
| | *rot*N | *rot*2 ('root') | 22 | 28 | **.81** | **.79** |
| | | *rot*7 ('mess') | 6 | | .19 | .21 |
| | *plan*N | *plan*1 ('scheme') | 35 | 43 | **.83** | **.81** |
| | | *plan*2 ('level') | 8 | | .17 | .19 |
| | *valg*N | *valg*1 ('choice') | 16 | 31 | .37 | **.52** |
| | | *valg*2 ('election') | 15 | | **.63** | .48 |
| | *slag*N | *slag*1 ('kind, type') | 19 | 51 | **.62** | .37 |
| | | *slag*2 ('battle') | 12 | | .16 | .24 |
| | | *slag*3 ('blow') | 18 | | .14 | **.35** |
| | | *slag*12 ('stroke') | 2 | | .09 | .04 |
| | *tak*N | *tak*2 ('grasp') | 161 | 226 | .11 | **.71** |
| | | *tak*4 ('ceiling') | 35 | | .37 | .15 |
| | | *tak*7 ('roof') | 30 | | **.52** | .13 |
| | *stemme*N | *stemme*1 ('voice') | 84 | 95 | **.94** | **.88** |
| | | *stemme*3 ('vote') | 11 | | .06 | .12 |
| | *liv*N | *liv*1 ('life') | 276 | 284 | **.98** | **.97** |
| | | *liv*12 ('waist') | 8 | | .02 | .03 |
| Adjectives | *full*AJ | *full*1 ('complete') | 316 | 324 | **.92** | **.98** |
| | | *full*2 ('drunk') | 8 | | .08 | .02 |
| | *gal*AJ | *gal*1 ('crazy') | 5 | 49 | .31 | .10 |
| | | *gal*2 ('incorrect') | 44 | | **.69** | **.90** |
| | *frisk*AJ | *frisk*1 ('fresh') | 27 | 53 | **.72** | **.51** |
| | | *frisk*4 ('healthy') | 26 | | .28 | .49 |
| Verbs | *trykke*V | *trykke*1 ('squeeze') | 22 | 26 | **.76** | **.85** |
| | | *trykke*2 ('print') | 4 | | .24 | .15 |
| | *utsette*V | *utsette*1 ('expose') | 57 | 69 | **.56** | **.83** |
| | | *utsette*3 ('postpone') | 12 | | .44 | .17 |

## 8.6 Conclusion

We conclude this chapter by a discussion of the representativity of the data sets. On the whole, the data set collection is good in that it covers all three open classes, and even though it was specified in (Section (8.2)) that we aim for target words

Table 8.19: 15 target words, **total sense-tagged data set** (automatically + manually annotated sets). The *a priori* sense probabilities (P(s) of each sense are given; the most frequent sense is marked in bold.

| | Lemma | Senses | Tags | Total instances | P(s) |
|---|---|---|---|---|---|
| Nouns | *lag*N | *lag*5 ('layer') | 36 | 54 | **66.7** |
| | | *lag*8 ('team') | 18 | | 33.3 |
| | *fyr*N | *fyr*1 ('guy') | 62 | 82 | **75.6** |
| | | *fyr*2 ('fire') | 20 | | 24.4 |
| | *utvalg*N | *utvalg*1 ('selection') | 30 | 66 | 45.5 |
| | | *utvalg*3 ('committee') | 36 | | **54.5** |
| | *rot*N | *rot*2 ('root') | 65 | 81 | **.80.2** |
| | | *rot*7 ('mess') | 16 | | .19.8 |
| | *plan*N | *plan*1 ('scheme') | 130 | 156 | **83.3** |
| | | *plan*2 ('level') | 26 | | 16.7 |
| | *valg*N | *valg*1 ('choice') | 60 | 149 | 40.3 |
| | | *valg*2 ('election') | 89 | | **59.7** |
| | *slag*N | *slag*1 ('kind, type') | 105 | 190 | **55.3** |
| | | *slag*2 ('battle') | 34 | | 17.9 |
| | | *slag*3 ('blow') | 37 | | 19.5 |
| | | *slag*12 ('stroke') | 14 | | 7.4 |
| | *tak*N | *tak*2 ('grasp') | 178 | 368 | **48.4** |
| | | *tak*4 ('ceiling') | 86 | | 23.4 |
| | | *tak*7 ('roof') | 104 | | 28.3 |
| | *stemme*N | *stemme*1 ('voice') | 444 | 478 | **92.9** |
| | | *stemme*3 ('vote') | 34 | | 7.1 |
| | *liv*N | *liv*1 ('life') | 1299 | 1324 | **98.1** |
| | | *liv*12 ('waist') | 25 | | 1.9 |
| Adjectives | *full*AJ | *full*1 ('complete') | 593 | 629 | **94.3** |
| | | *full*2 ('drunk') | 36 | | 5.7 |
| | *gal*AJ | *gal*1 ('crazy') | 41 | 167 | 24.6 |
| | | *gal*2 ('incorrect') | 126 | | **.75.4** |
| | *frisk*AJ | *frisk*1 ('fresh') | 74 | 118 | **62.7** |
| | | *frisk*4 ('healthy') | 44 | | 37.3 |
| Verbs | *trykke*V | *trykke*1 ('squeeze') | 52 | 67 | **77.6** |
| | | *trykke*2 ('print') | 15 | | 22.4 |
| | *utsette*V | *utsette*1 ('expose') | 75 | 110 | **68.2** |
| | | *utsette*3 ('postpone') | 35 | | 31.8 |

that are as uncontroversial as possible, it seems that some of them may be quite challenging after all.

As for the data set sizes, they are generally low by the ideal lexical sample standards (cf. Chapter (3.4.2), p. 38). The total data set of the presented Norwegian lexical sample (4039 instances totally, divided across 15 target words) has on average $269 \pm 337$ corpus instances, the minimum number being 54 instances and the maximum being 1324 (based on the counts in Table (8.19)). Dividing each of these data sets into 70% as the development set and 30% as the held-out data set, there is on average 188 examples for training and 80 for testing per target word (the minimum and maximum number of training instances is 37 and 926, respectively; the minimum and maximum number of test instances is 16 and 397, respectively).

By comparison, the Swedish lexical sample introduced in SENSEVAL-2 contained 40 lemmas (20 nouns, 15 verbs and 5 adjectives), totalling 8718 training/development instances and 1525 test instances (Lager & Zinovjeva, 2001), which on average yields 218 training instances and 38 test instances per word. Specia et al. (2009) use data from the SEMEVAL-2007 English lexical sample task, using corpus examples of 65 verbs and 35 nouns. Their data sets have an average of 222 examples for training and 49 for testing per target word (the minimum no. of training instances is 19 training examples and 2 for test material, the maximum is 2,536 training instances and 541 test instances).

This suggests although on the low side, the data set for Norwegian words are within the limits of what is acceptable in the WSD community. The collection of more material from another resource than the ENPC is beyond the scope of the present project, since the motivating factor for the entire project concerns whether one may 'enlarge' the information value of small data sets by adding Mirrors-derived information from about the words surrounding an ambiguous target word. The systematic experiments for testing with and without Mirrors-information about context words makes the current thesis bound to the ENPC, since it is word-aligned (and since it was chosen to experiment with Norwegian material).

A note on the sense distribution will be made at the end. Agirre and Martínez (2000, 2004) find that the sense distribution constitutes a statistical bias: WSD performance may degrade significantly if the training and test data have a different sense distribution. This is logical, since the computed sense probabilities from the training set will then not fit the test data. In other words, their findings suggest that the senses do not need to have an even distribution, but it is unfortunate if the sense distribution is not approximately equal in the training material and in the test sample. We therefore expect that although the sense distribution is sometimes skewed, this need not represent a problem for the classification experiments. The held-out data set is a set of randomly chosen instances from the total set, so no at-

tempt was made to ensure a perfectly equal sense distribution in the development sets and in the held-out test sets.

# CHAPTER 9

# COMPARING AND COMBINING KNOWLEDGE SOURCES

## 9.1 Introduction

This chapter presents the set of experiments stated in Chapter (6.4.1), p. 120. Four experiments are conducted for each target word, in which different knowledge sources are applied using otherwise identical experimental settings. The set of experiments (abbreviated EXP) is as follows:

| | |
|---|---|
| **EXP1** | The [$\pm$ $n$] nearest WORDS (Ws). |
| **EXP2** | The SEMANTIC-FEATURES (SFs) derived from those words in EXP1 that were automatically sense-tagged. |
| **EXP3** | The RELATED-WORDS (REL-W) derived from those words in EXP1 that were automatically sense-tagged. |
| **EXP4** | Combined classifier W + SF + REL-W (the most confident gets to vote). |

As stated in Chapter (6.4.1), these experiments address the following research questions:

- How well may a traditional WORD classifier (EXP1) be expected to perform, given our specific data sample, sense inventory and classification algorithm?
- *Replacing* context words with Mirrors-derived information (EXP2, EXP3): what is the loss or gain in using Mirrors-derived information as individual resources with respect to EXP1?

- *Adding* Mirrors-derived information (EXP4): What is the loss or gain in adding paradigmatic information from the Mirrors method? Does added information in fact lead to more confident and more correct classifications? (or does Mirrors-derived information, quite to the contrary, introduce more noise?)

In Section (9.2) we follow Pedersen (2000) in conducting model selection by systematic tests of combinations of context window sizes. The best classification model is then tested on a held-out test set in Section (9.3).

## 9.2 Model selection

### 9.2.1 Model selection setup

In line with the acknowledgement of WSD as a 'word expert' task (Chapter (3.2)), where not all words benefit from the same kinds of contextual information, it is common in data-based WSD experiments to conduct a series of development experiments. The purpose of these experiments is to determine how to *instantiate* the model, by considering what seems to be the best model settings based on the development data material. Given the total data set available for a target word, $30\%$ of the instances are picked randomly and allotted for held-out evaluation (the same held-out test set is used for all held-out experiments), whereas the remaining $70\%$ are used as the development data set.

For model selection, the approach described in Pedersen (2000) is adopted.

#### The original approach in Pedersen (2000)

Pedersen builds a so-called *ensemble* of $9$ individual classifiers that represent different views of the contextual characteristics of a target word in virtue of having learnt from different context information. For each given instance of the target word, each of the nine classifiers gives its vote for the most probable sense, and the sense that receives the majority of votes is assigned to the instance.

Note that we only adopt the preliminary step of *model selection* as described in Pedersen (2000), in which $9 \times 9$ classifiers are tested, as it provides an appreciably clear way to test many classifiers systematically. The basic skeleton of his approach is illustrated in Table (9.1) (p. 199). Whereas Pedersen (2000) applies a model selection framework to then select the *nine* best classifiers to be used in the final ensemble framework, we only choose the best classifier which is then used in the final classification experiments on held-out test data. After first sketching the original model selection approach in Pedersen (2000), we will then consider how the approach is adapted to suit the needs of the current dissertation.

Two parameters are varied in the model selection approach, viz. the left-side ($l$) and the right-side ($r$) context window. The combinations of these parameters are thus conveniently visualised by a table where the vertical axis contains the $l$ values and the horizontal axis contains the $r$-values. Both context windows are given nine possible sizes, so all combinations of $l$ and $r$ yield $81$ parameter combinations totally. For each combination a Naive Bayes classifier is trained, notated as $naivebayes(l, r)$. For instance, if the left-side is set to $4$ and the right-side to 2, this classifier is notated as $naivebayes(4, 2)$. Each classifier is evaluated with $5$-fold cross-validation. This means that the training material is randomly partitioned into $5$ folds, where each fold at a time is used once as test material while the other four partitions serve as training material.

Table 9.1: Skeleton of the $9 \times 9$ classifiers in model selection.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| wide | 50 | | | | | | | | | |
| | 25 | | | | | | | | | |
| | 10 | | | | | | | | | |
| medium | 5 | | | | | | | | | |
| | 4 | | | | | | | | | |
| | 3 | | | | | | | | | |
| narrow | 2 | | | | | | | | | |
| | 1 | | | | | | | | | |
| | 0 | | | | | | | | | |
| | | 0 | 1 | 2 | 3 | 4 | 5 | 10 | 25 | 100 |
| | | narrow | | | medium | | | wide | | |

Pedersen (2000) aims to avoid that the ensemble consists of classifiers with similar context window sizes, since the 'ultimate success of an ensemble depends on the ability to select classifiers that make complementary errors' (Pedersen, 2000, p. 68). Therefore, the nine possible context window values are categorised into groups of similar context window sizes, and only the best classifier in each such category is selected to vote in the final ensemble system. The three lowest window sizes are categorised as *narrow*, the three intermediate window sizes as *medium* and the three widest window sizes as *wide*. Combining a left-side and a right-side window category, nine possible combinatorial categories emerge: narrow–narrow, narrow–medium, narrow-wide, wide–narrow, etc. Pedersen then selects the best classifier in each such category, thus obtaining nine classifiers to be used in his final ensemble classification system. Since our model selection, by contrast, only focusses on the best classifier (and not the nine best), the nine combinatorial categories do not serve a *methodological* purpose. But since it is natural to expect smaller differences among the classifiers in one category, the size categorisation (narrow, medium and wide) is still helpful to get an overview over the range of classifiers.

**Adapting the model selection framework**

For the present thesis, Pedersen (2000)'s model selection approach is modified in three ways: First, whereas Pedersen (2000) includes open-class and closed-class co-occurrences, we restrict our attention to open-class co-occurrences only. The reasons for this are discussed in Chapter (6.3.1) (p. 108); basically it is a methodological point to focus on the kind of words that can be related to Mirrors-derived classes of semantically related words.

Second, the nine parameter values differ from those of Pedersen (2000). Pedersen suggests the following nine values to the left-side and right-side context window which are in the range of zero (no context at all) and $50$:

$$[0, 1, 2, 3, 4, 5, 10, 25, 50]$$

We introduce, firstly, a new set of context window sizes when training on WORDS and SEMANTIC-FEATURES, and secondly, a new kind of parameter for the RELATED-WORDS, viz. the *SynsetLimit*. The proposed window context sizes of the current dissertation is given below. Compared with Pedersen's values, the zero is omitted and the maximum window size is adjusted upwards; correspondingly the intermediate values are also adjusted to suit the new range limits. The left-side and right-side context window values used in the current thesis are:

$$[1, 2, 4, 10, 20, 30, 50, 75, 100]$$

The reason for omitting the zero value is that this is in practice a most frequent sense (MFS) classifier. We use MFS as a baseline, against which the context-based classifiers are compared, therefore the MFS classifier is not included *in* the experiments. The maximum context window is adjusted up from a maximum of $50$ to a maximum of $100$, since Yarowsky and Florian (2002) find that as many as $\pm 150$ context features may be useful (at least for nouns) with a Naive Bayes model.

When training on the RELATED-WORDS a new kind of experimental parameter is introduced, viz. the *SynsetLimit*. It is introduced in accordance with the observation in Chapter (6) that it might be of interest to experiment with differing SynsetLimits. In order to maintain the experimental format of cross-combining $9 \times 9$ classifiers, the first parameter (the vertical dimension in Table (9.1)) is constituted by *the nine best context window sizes from the WORD classifiers*, WORDS being the 'best-known' approach among the approaches tested here. The second parameter (the horizontal dimension) is the Synset Limit. The SynsetLimit values used in the current thesis are the following:

$$[automatic, 4, 5, 10, 15, 20, 30, 40, 50]$$

The automatic value is computed by Equation (4.1) in Chapter (4), and its precise value will vary according to the semantic field within which a word sense occurs. The range limits of the fixed SynsetLimit values were set in discussions with Helge Dyvik. The values have not previously been subject to systematic studies, but in Dyvik's general experience the value $4$ seems to represent a minimum value, since values below this usually does not lead to any changes in the way word senses are grouped together in terms of relatedness. The value $20$ has generally been observed to represent a useful maximum. It often seems to be the case that a higher SynsetLimit value causes too many word senses to be included, although this impression has never been quantified in any way. It was thus considered worthwhile to test whether this seeming tendency is reflected in quantitative results.

Third, and finally, the current presentation deviates from Pedersen (2000) in how the quantitative results are presented. Pedersen uses the average recall of the five cross-validation folds. In the model selection work of the present dissertation the standard deviation between each cross validation fold was not as consistent as in Pedersen's study because of smaller data sets. Therefore, overall accuracy is used instead, as used for instance in the TiMBL software package (Daelemans et al., 2007) and in Specia et al. (2009)[1]. Overall accuracy (OA) is computed as the ratio of correct classifications and instances to be classified.

It may also be commented that as specified in the experimental outline (Chapter (6)) it was originally a point that the implemented classifier does not back off to the most frequent sense if it nothing in the context is known from training. Originally it was then intended to be a point to study whether the Mirrors-derived knowledge sources will have learnt more context words than a simple WORD model and, if so, whether this would motivate studies on the relations between precision and recall for a classifier. As it turned out in the classified development material, however, it simply did not occur very often that instances are left unclassified, consequently the precision and recall measures do not diverge much. Therefore we will only discuss the recall measures, and instead simply mention where relevant when classifiers could not classify some test instances.

So, with five-fold cross-validation the overall accuracy is based on the counts from each test fold. For instance, the $naivebayes(4, 4)$ classifier for *utvalg*N with a simple WORDS classifier yielded the following 5-fold results:

[1] Specia et al. (2009) call their evaluation measure an averaged accuracy based on five-fold cross-validation, but the term 'average' is not accurate since they simply sum the counts from each cross validation fold, as we also do in the presented experiments of the present thesis.

((8 . 10) (9 . 9) (9 . 9)(7 . 9)(6 . 9))

(i.e. 8 correct classifications of 10 instances to be classified in fold 1, 9 correct of 9 in fold 2, etc.). This yields an overall accuracy (OA) of

$$\text{OA}=\frac{8+9+9+7+6}{10+9+9+9+9} = \frac{39}{46} = .848$$

In other words, the cross-validation tests for *utvalg*N, with a word-co-occurrence (WORDS) model set to $naivebayes(4, 4)$, had an overall accuracy of $84.8\%$ (the baseline, computed as the relative frequency of the most frequent sense in the total development set: $60.9\%$).

The development classifiers are eventually ranked in terms of performance in order to choose the best-performing one for held-out evaluation. In case of ties we follow Pedersen (2000) and select the classifier with the smallest total context window size.

### 9.2.2 Results (cross-validation)

#### General analysis of the cross validation results

The complete tables for the model selection experiments of each target word are listed (alphabetically) in Appendices 3, 4 and 5. Appendix 3 shows the cross validation results for WORDS (EXP1), Appendix 4 for SEMANTIC-FEATURES (EXP2) and Appendix 5 for RELATED-WORDS (EXP3). The appendices list one table for each target word and each table shows the overall accuracy (recall) for each of the 81 classifiers. The baseline, computed as the relative frequency of the *most frequent sense* (MFS) in the total development data set, is given above each table. For instance, the best classifier of *frisk*AJ (the alphabetically first listed target word) in Appendix 3 was $naivebayes(4, 2)$ (a narrow-narrow window), which had an overall accuracy (measured as recall) of $69.5\%$. The MFS (most frequent sense) baseline was $68.3\%$. The best classifier in each of the nine window size categories is marked in bold.

For the sake of overview, the best classifier for each target word from cross-validation is summed up in three tables below: WORDS in Table (9.2) (p. 203), SEMANTIC-FEATURES in Table (9.3) (p. 204) and RELATED-WORDS in Table (9.4) (p. 204). The first column lists the target words (abbreviated TW). The next columns list the MFS (the baseline), the best $naivebayes(l, c)$ (context window size) setting for the given target word, the window size category to which this best setting belongs and Overall Accuracy. Accordingly, in the appendix example mentioned above, *frisk*AJ, is listed with a baseline of $68.3$, its best context window

setting was $naivebayes(4, 2)$, the abbreviation *NN* denotes a *narrow-narrow* window (M and W denotes medium and wide, respectively), and the overall accuracy based on 5-fold cross-validation was $69.5\%$. It may be noted that in the case of RELATED-WORDS, the *narrow–medium–wide* categorisation is a bit misleading, since its nine values on the horizontal axis of the tables are the nine best models from the WORD experiments (so for instance, one of the values in the RELATED-WORDS table for *frisk*AJ in Appendix 5 will be $(2, 4)$, since this was the best model from the WORDS model selection. Therefore, this categorisation has not been applied to the reported RELATED-WORDS experiments.

Based on the detailed results in Appendices 3–5, and as also indicated by the best classifiers listed in tables (9.2—9.4), the best performing classifiers are generally found using window size values in the medium or narrow context size window category or in combinations of these (i.e. using context window sizes in the range of $1$ and $30$).

Table 9.2: The best classifier from cross-validation: WORDS

| TW | MFS(%) | $naivebayes(l, r)$ | category | OA (%) |
|---|---|---|---|---|
| *frisk*AJ | 68.3 | (4,2) | *NN* | 69.5 |
| *full*AJ | 94.1 | (2,4) | *NN* | 86.6 |
| *fyr*N | 78.9 | (30,20) | *MM* | 89.5 |
| *gal*AJ | 77.6 | (10,1) | *MN* | 77.6 |
| *lag*N | 70.3 | (2,20) | *NM* | 97.3 |
| *liv*N | 98.1 | (2,2) | *NN* | 95.0 |
| *plan*N | 87.2 | (2,4) | *NN* | 87.2 |
| *rot*N | 80.4 | (2,30) | *NM* | 89.3 |
| *slag*N | 55.6 | (1,10) | *NM* | 57.9 |
| *stemme*N | 92.2 | (10,10) | *MM* | 99.1 |
| *tak*N | 47.5 | (2,2) | *NN* | 67.7 |
| *trykke*V | 80.4 | (10,4) | *MN* | 89.1 |
| *utsette*V | 67.5 | (30,4) | *MN* | 83.1 |
| *utvalg*N | 60.9 | (30,2) | *MN* | 95.7 |
| *valg*N | 60.6 | (4,30) | *NM* | 93.3 |

**WORDS development results**

Considering each knowledge source in order and beginning with WORD models, Table (9.2) (p. 203) shows that the narrow and medium window sizes clearly dominate. The same tendency is seen when scrutinising the full tables in Appendix 3; the best classifiers of each target word are generally clustered among combinations with medium and narrow window sizes (although the extremely narrow windows tend to fare less well).

Viewing the WORDS model as a kind of 'baseline' representing how well a classifier could be expected to perform with the Mirrors sense inventory and the given data material, the 'learnability' of the Mirrors senses of the target words

Table 9.3: The best classifier from cross-validation: SEMANTIC-FEATURES

| TW | MFS(%) | $naivebayes(l, r)$ | category | OA (%) |
|---|---|---|---|---|
| *frisk*AJ | 68.3 | (4,2) | *NN* | 62.2 |
| *full*AJ | 94.1 | (2,1) | *NN* | 73.9 |
| *fyr*N | 78.9 | (4,2) | *NN* | 64.9 |
| *gal*AJ | 77.60 | (1,2) | *NN* | 74.1 |
| *lag*N | 70.3 | (100,75) | *WW* | 75.7 |
| *liv*N | 98.1 | (2,2) | *NN* | 92.5 |
| *plan*N | 87.2 | (2,2) | *NN* | 80.7 |
| *rot*N | 80.4 | (1,4) | *NN* | 76.8 |
| *slag*N | 55.6 | (1,4) | *NN* | 45.9 |
| *stemme*N | 92.2 | (4,1) | *NN* | 96.4 |
| *tak*N | 47.5 | (1,2) | *NN* | 58.0 |
| *trykke*V | 80.4 | (50,75) | *WW* | 71.7 |
| *utsette*V | 67.5 | (4,2) | *NN* | 68.8 |
| *utvalg*N | 60.9 | (20,1) | *MN* | 78.3 |
| *valg*N | 60.6 | (10,2) | *MN* | 73.1 |

Table 9.4: The best classifier from cross-validation: RELATED-WORDS

| TW | MFS(%) | $naivebayes(l, r)$ | OA (%) |
|---|---|---|---|
| *frisk*AJ | 68.3 | ((2,30) , 15) | 70.7 |
| *full*AJ | 94.1 | ((4,1) , 40) | 86.1 |
| *fyr*N | 78.9 | ((30,10) , 10) | 84.2 |
| *gal*AJ | 77.6 | ((1,2) , 50) | 75.9 |
| *lag*N | 70.3 | ((30,10) , 4) | 94.6 |
| *liv*N | 98.1 | ((1,2) , 50) | 94.5 |
| *plan*N | 87.2 | ((10,2) , 20) | 84.4 |
| *rot*N | 80.4 | ((4,30) , 20) | 85.7 |
| *slag*N | 55.6 | ((30,4) , 40) | 57.9 |
| *stemme*N | 92.2 | ((10,10) , 10) | 99.4 |
| *tak*N | 47.5 | ((2,2) , 10) | 66.9 |
| *trykke*V | 80.4 | ((20,30) , 15) | 87.0 |
| *utsette*V | 67.5 | ((10,20) , 30) | 77.9 |
| *utvalg*N | 60.9 | ((10,30) , 15) | 95.7 |
| *valg*N | 60.6 | ((20,10) , 30) | 90.4 |

seems promising in that all target words had classifiers that performed above or equal to baseline, except *full*AJ (baseline: $94.5\%$) and *liv*N (baseline: $98.4\%$). Both with *full*AJ and *liv*N the baselines are so extremely high that the failure to learn the baseline is perhaps not very surprising. But as illustrated by the lemma *stemme*N, a high baseline is not impossible to beat: *stemme*N has almost as high baseline as *full*AJ (baseline: $93.4\%$), and its best classifier exceeds this baseline by $6.9$ points ($99.1$). The remaining thirteen target words had between one and sixty-six classifiers above the baseline. For the sake of overview, these target words are listed on the next line, ordered according to their number of classifiers above the baseline (the number of classifiers is listed in parenthesis):

**Number of classifiers above baseline in the development phase (WORDS):** *lag*N (66), *utvalg*N (65), *valg*N (38), *utsette*V (36), *tak*N (33), *stemme*N (27), *rot*N (26), *fyr*N (20), *trykke*V (19), *frisk*AJ (5), *plan*N (3), *slag*N (3), *gal*AJ (1).

### SEMANTIC-FEATURES development results

Moving on to the cross-validation results for SEMANTIC-FEATURES, two observations are noted. First, surprisingly few target words had any classifiers above the baseline. Whereas the WORDS model had thirteen target words above or equal to the baseline, only six target words met or beat the baseline based on SEMANTIC-FEATURES. The six target words with at least one model above the baseline are listed on the next line in order of their number of classifiers above the baseline (the exact number of classifiers is listed in parenthesis):

**Number of classifiers above baseline in the development phase (SEMANTIC-FEATURES):** *utvalg*N (27), *lag*N (14), *tak*N (14), *valg*N (11), *stemme*N (7), *utsette*V (1).

This observation clearly indicates that the classification task became harder when based on SFs. A possible explanation for this could be the information loss caused by restricting our attention to context words that were sense-tagged—if so, this potential factor is eliminated in Chapter (10), in which the WORDS classifiers only consider the $n$ nearest words that were sense-tagged; thus one can more directly measure the loss or gain in the usefulness of Mirrors-derived information compared to the lemmas in the WORDS model. But if this was alone the explanation for the seeming weaker results with a SF model, one should expect the RELATED-WORD experiments (Table (9.4)) to be at the same level as the SEMANTIC-FEATURE experiments; this is however not the case.

Therefore, the classification output with a WORD model was compared with the corresponding SEMANTIC-FEATURE model in order to study which context information that contributed to more correct classifications in the WORD model than in the SEMANTIC-FEATURE model. The comparison suggests that the frequency counts for a context feature are often higher in the SF model, but in return these

generalised features do not seem to pull equally clearly in the direction of one particular target word sense. In other words, it seems that the SFs may be too general. An example from the cross validation output will illustrate this.

The target word *rot*N had 26 classifiers above the baseline with the W model and none when replacing Ws by SFs, so this is one of the target words with a fairly clear, general performance decline from Ws to SFs. The following example is taken from the best W model for *rot*N, $naivebayes(2, 30)$ (cf. Table (9.2) (p. 203)). This model was compared with the corresponding SF model, which means that exactly the same sets of context words form the basis for learning, but in the SF model sense-tagged context words are replaced by their Mirrors-derived semantic features whereas context words that were not sense-tagged are simply 'lost'. Example (13) shows the immediate context of a target word instance that was correctly classified with the W model and wrongly classified with the SF model. As can be seen from the immediate context, this instance represents the 'root' sense of *rot*N—henceforth ROOT—(as opposed to the 'mess' sense—henceforth MESS):

(13) (. . .) å vende tilbake til de andre **røttene**|**ROOT** sine også.[2]
(. . .) to return to his other roots as well.

Now, every context feature that is known from training may be termed a *contributing* feature because for every known context feature, the probability of seeing this context feature given a target word sense is computed. As specified in Chapter (6.5.2) (p. 126 and onwards), this is computed by the MLE (maximum likelihood estimation) in Equation (6.4). The exact contribution of each context feature to the final classification outcome may be ranked by considering the ratio between its MLE per target word sense: the higher the ratio, the more it 'pulls' in the direction of a given target word sense. Importantly, this applies even if the context feature is lower-frequent.

Ranking the lemmas that contributed to the correct classification of Example (13) in the W model, four context features (lemmas) shared the top rank. Table (9.5) lists the five most contributing context features. The significant point of this table is that the fifth most contributing context feature, *kunne*V, co-occurred more times with the correct sense (16 times) than the four highest-ranked features (which occurred two times each), and has a higher probability (a higher MLE value) than the four highest-ranked features.

Still, the four top-ranked context features have a higher ratio between their MLE s given ROOT as opposed to MESS (the ratio is $0.8369561$[3]). By contrast, the

[2]The sentence ID in the ENPC is: ABR1T

[3]Recall from Chapter (6.5.2), when computing this, that in order to avoid zero counts, every frequency is incremented by $0.1$.

Table 9.5: The five most contributing context features with the WORDS model with respect to Example (13)

| context feature | ROOT | | MESS | |
|---|---|---|---|---|
| | $N$ | MLE | $N$ | MLE |
| *reise*V ('travel') | 2 | 0.046667 | 0 | 0.009091 |
| *spesiell*AJ ('special') | 2 | 0.046667 | 0 | 0.009091 |
| *samfunn*N ('society') | 2 | 0.046667 | 0 | 0.009091 |
| *hode*N ('head') | 2 | 0.046667 | 0 | 0.009091 |
| *kunne*V ('could') | 16 | 0.357778 | 1 | 0.1 |

next-ranked lemma in the WORD model, *kunne*V had a resulting ratio between the MLEs of $0.7815535$. This example illustrates that the most contributing context feature is not necessarily the one with the highest probability; the important issue is the *ratio* between likelihoods.

We may now consider the corresponding SEMANTIC-FEATURE model, where Example (13) was classified erroneously. First, the most contributing feature was a semantic feature that emerged from the verb lemma *vende*V ('(re)turn'), a lemma that was actually unknown in the WORD model (i.e. it was never really attested in the training corpus for *rot*N). This context feature co-occurred four times with MESS and never with the correct sense, ROOT, in the training material. The second most contributing context feature was a semantic feature that emerged from the verb lemma *reise*V ('travel'), which was the context lemma that contributed most to the correct classification in the WORDS model. But in the SF model the SEMANTIC-FEATURE associated to *reise*V was registered once together with MESS and never with ROOT, i.e. the frequency is not only lower than in the W model but it also pulls in the opposite direction. As a third example, another of the four most strongly contributing context lemmas in the WORD model was *spesiell*AJ ('special'), which in the WORD model co-occurred twice with ROOT and never with MESS. In the SF model, one of its semantic features was seen only once with ROOT and never with MESS.

In other words, we see that the frequencies change and even pull in opposite directions when abstracting from WORDS to the SEMANTIC-FEATURES of these context words. When frequency counts are reduced, as with the *spesiell*AJ example just mentioned, this may in itself be explained because of the expected information loss that may arise when abstracting to the subset of context words that have semantic information during training. As for the observation that observed context features may pull in contradictory sense directions when comparing the WORD and SEMANTIC-FEATURE models, this may suggest that the SEMANTIC-FEATURES are too general (or simply not plausible); this will be further analysed when moving on to the held-out data sets.

A second thing to note about SEMANTIC-FEATURES is that the narrow context windows clearly dominate; in all but two cases the best classifier is composed of two narrow windows (i.e. a *narrow-narrow* model). Based on the full model tables in Appendix 4 this is a general pattern in that the better-performing classifiers of a target word (and not just the single best classifier) are clustered in and around the narrow window category. Two target words form exceptions to this in having wide context windows as their best model setting; *lag*N ($naivebayes(100, 75)$) and *trykke*V ($naivebayes(50, 75)$). Looking at the full tables for these two target words, it does not seem that the model selection could find a clearly superior context window category at all. In the case of *lag*N, only half of its top ten models belong the *wide-wide* (*WW*) category; the others belong to the medium-narrow and narrow-narrow window category. Likewise, with *trykke*V, its ten best classifiers belong to a variety of window categories (medium-narrow windows, wide-medium, narrow-medium, wide-medium and wide-wide windows).

#### RELATED-WORDS development results

Moving on to RELATED-WORDS, the performance level improves quite markedly compared to SEMANTIC-FEATURES: eleven of the fifteen target words have classifiers equal to or exceeding baseline. Two of these did not reach the baseline with WORDS (or, for that matter, SEMANTIC-FEATURES), either, viz. *full*AJ and *liv*V; the two other target words that failed to meet or beat the baseline is *gal*AJ (baseline: $77.6$) and *plan*N (baseline: $87.2$). The target words with at least one model above the baseline are listed on the next line in order of their number of classifiers above the baseline (the exact number of classifiers is listed in parenthesis):

**Number of classifiers above baseline in the development phase (RELATED-WORDS):** *lag*N (81), *tak*N (81), *utvalg*N (81), *valg*N (81), *stemme*N (74), *utsette*V (65), *trykke*V (35), *rot*N (28), *fyr*V (23), *frisk*AJ (12), *slag*N (3).

Most of the target words have a slightly lower overall accuracy in the RELATED-WORDS experiments for their best classifier than the corresponding best classifiers for WORDS. In other words, the performance of the RELATED-WORDS classifiers were generally quite good, although usually slightly below the corresponding WORDS classifiers. At the same time, the RELATED-WORDS classifiers in the model selection experiments are more stable with regard to the parameter variation, compared to WORDS and SEMANTIC-FEATURES. Recall that the RELATED-WORDS experiments differ from the experiments with WORDS and SEMANTIC-FEATURES in that instead of testing $9 \times 9$ context window sizes, only the nine best window sizes from the WORDS experiments are used in combination with nine different SynsetLimit values. Hence, the presumably least successful context window sizes are already weeded out in the RELATED-WORD experiments.

From the development experiments there does not seem to be a clear tendency with regard to the best SynsetLimit value; it may be remarked that because of this we ran a test based on an automatically set SynsetLimit the RELATED-WORD model selection with the left-side and right-side windows as the two variable parameters. These model selection results proved, however, to produce slightly lower results when setting the SynsetLimit automatically by default, therefore it seems worthwile to keep the SynsetLimit as a variable parameter.

**Summing up the development results**

In sum, the general tendency from the development results is that the WORD-based classifiers seem to perform best in terms of overall accuracy, the RELATED-WORDS come quite close to WORDS whereas the SEMANTIC-FEATURES-based results seem less promising. The next section presents the results when training on the total development set using the best development classifier for each target word, testing on the held-out test set. Since we do not really expect Mirrors-derived information to compete with a WORD-based model in these experiments, the Mirrors-derived knowledge sources being restricted to the subset of word co-occurrences that were sense-tagged in context, we will pay particular attention to the extent to which Mirrors-derived information may have complementary qualities with respect to a simple, traditional WORD model.

## 9.3 Evaluating on held-out data sets

In this section we present the result of applying each of the three knowledge sources in isolation (WORDS, SEMANTIC-FEATURES and RELATED-WORDS) on the same held-out data sets.

Based on the model selection in the preceding section, the best model of each target word from the model selection experiments is tested on the held-out data set. As specified in Chapter (8), the allotted $30\%$ (randomly selected) yields on average $80$ instances for testing per target word (the minimum and maximum number of test instances being $16$ and $397$, respectively). It is sometimes convenient to sum the classification outcomes across all fifteen target words to obtain a simplified outcome per knowledge source (among other things this is convenient to say something about the statistical significance of differences between the knowledge sources); in all the sum of test instances is $1219$.

The classification outcomes will be analysed from three points of view:

- Comparing each knowledge source against the baseline (the most frequent sense, MFS).

- Comparing the two Mirrors-derived knowledge sources against the 'traditional' WORDS classifier, the latter representing the 'best-known' approach.

- Analysing whether the classifiers make complementary errors.

In Section (9.3.2)—Section (9.3.5) the individual classifiers are evaluated. Then, the three knowledge sources are combined in a voting scheme where the most confident gets to vote (Section (9.3.6)).

### 9.3.1 Some basic terminology

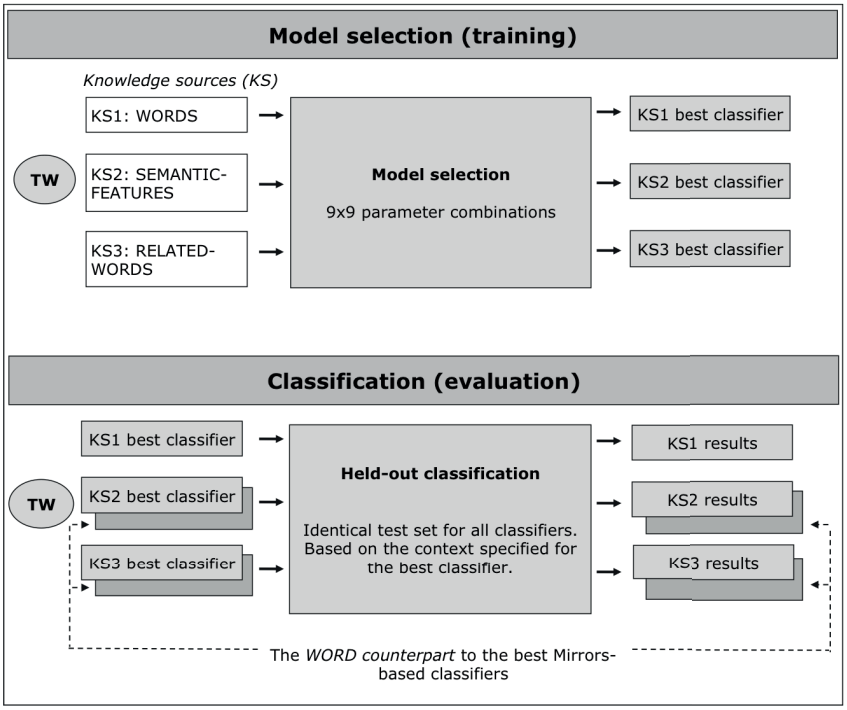
So far we have discussed how the model selection phase serves to select the best classifier settings for each of the three knowledge sources that we use for each target word (see the top half section of Figure (9.1) (p. 211)). Having now arrived at the classification evaluation, the time is ripe to introduce the concept of WORD models that constitute the *direct counterparts* to the best classifier settings used by a Mirrors-based knowledge source (see the lower half section of Figure (9.1) (p. 211)).

Testing the three knowledge sources (abbreviated KS in the figure) of a target word TW individually amounts to testing each knowledge source with its *best* classifier from the model selection phase. This implies that the different knowledge sources may depend on entirely different context window sizes. On the one hand, this means that each knowledge source should have a chance to perform maximally well, based on the optimal parameter combination according to the model selection phase. On the other hand, this also means that the three classifiers cannot be directly juxtaposed when it comes to analysing the effect of abstracting from words to Mirrors-derived information.

Say for instance that the best WORDS model is the $naivebayes(20, 20)$ (i.e. collecting the $20$ nearest lemmas on each side of the target word during training and, subsequently, during classification) whereas the best SEMANTIC-FEATURES model for the same target word is $naivebayes(2, 2)$. In that case the SEMANTIC-FEATURES model collects all semantic features that can be retrieved based on the two nearest lemmas (to the extent that lemmas within this context window were sense-tagged) on each side of the target word. With such different settings it remains interesting if one of the knowledge sources is clearly better than the other. As regards the level of abstraction with a Mirrors-based knowledge source, however, it must be based on the available lemmas within same context window from which Mirrors-based knowledge was retrieved.

Therefore, a so-called 'WORD counterpart' to a Mirrors-based classifier is a WORD model which is trained and evaluated with the same context window settings as that used by the Mirrors-based classifier in question. The role of the

Figure 9.1: Model selection and evaluation on the best classifiers from model selection and on the direct WORD counterparts to the best SEMANTIC-FEATURES and RELATED-WORDS




WORD counterpart is illustrated in Figure (9.1) (p. 211) by partially 'hiding' them behind the best KS2 or KS3, thus showing that the training and classification results of the WORD counterparts are only relevant with respect to the Mirrors-derived classifier.

### 9.3.2 Evaluating the knowledge sources individually

#### Overview

The results when testing each classifier on the same data sets are summed up in Table (9.6) (p. 212). The target words are listed in the first column and the second

column lists the baseline, i.e. the number of test instances that would be correctly sense-tagged when always assigning the most frequent sense (the MFS) from training. The final three columns show the overall accuracy when applying each knowledge source individually. Each model is evaluated in terms of overall accuracy; the table shows the overall accuracy measured in absolute numbers (e.g. $20/36$ in the first row of the WORD model means $20$ correct classifications out of $36$ test instances to be classified) and with the corresponding ratio (e.g. $20/36 = .556$). Below the fifteen target word results in the table is the sum of correct classifications totally per knowledge source across all target words.

Table 9.6: Overall Accuracy for the individual knowledge sources (#=absolute counts, %=relative proportion

| TW | MFS | WORDS | | SEMANTIC-FEATURES | | RELATED-WORDS | |
|---|---|---|---|---|---|---|---|
| | | # | % | # | % | # | % |
| *frisk*AJ | .500 | 20/36 | .556 | 22/36 | .611 | 29/36 | .806 |
| *full*AJ | .947 | 156/189 | .825 | 140/189 | .741 | 165/189 | .873 |
| *fyr*N | .680 | 18/25 | .720 | 18/25 | .720 | 19/25 | .760 |
| *gal*AJ | .706 | 33/51 | .647 | 34/51 | .667 | 34/51 | .667 |
| *lag*N | .647 | 12/17 | .706 | 13/17 | .765 | 13/17 | .765 |
| *liv*N | .982 | 383/398 | .962 | 372/398 | .935 | 376/398 | .945 |
| *plan*N | .745 | 38/47 | .809 | 32/47 | .681 | 37/47 | .787 |
| *rot*N | .800 | 16/25 | .640 | 19/25 | .760 | 18/25 | .720 |
| *slag*N | .544 | 34/57 | .596 | 29/57 | .509 | 35/57 | .614 |
| *stemme*N | .944 | 144/144 | 1.000 | 135/144 | .937 | 142/144 | .986 |
| *tak*N | .505 | 83/111 | .748 | 74/111 | .667 | 78/111 | .703 |
| *trykke*V | .714 | 16/21 | .762 | 11/21 | .524 | 15/21 | .714 |
| *utsette*V | .697 | 27/33 | .818 | 22/33 | .667 | 26/33 | .788 |
| *utvalg*N | .550 | 20/20 | 1.00 | 17/20 | .850 | 20/20 | 1.000 |
| *valg*N | .578 | 38/45 | .844 | 36/45 | .800 | 33/45 | .733 |
| Total | | 1038/1219 | .852 | 974/1219 | .799 | 1040/1219 | .853 |

The experiments of the current chapter start from the $n$ nearest lemmas surrounding a target words, regardless if these lemmas have been sense-tagged with Mirrors senses. It would then not be surprising if the two Mirrors-based knowledge sources performed markedly worse than the traditional WORD-based classifier. However, the overall best classifier, based on the total number of correct classifications (the *Total* row below the results per classifier), is actually the RELATED-WORDS. This is only a marginal lead over WORDS, though, being better than the WORD-based classifier with a ratio of $.853$ against $.852$. This difference is not statistically significant as judged by McNemar's test with $p$=$0.05$ (cf. Chapter (6.6.3), p. 133).

Considering the fifteen target word outcomes in detail, the WORD classifier has the highest accuracy (or the same value as the winning accuracy) with eight out of the fifteen target words; the RELATED-WORDS with seven of the target words. The SEMANTIC-FEATURES classifier is only best with three of the fifteen target

words and the RELATED-WORDS classifier is thus clearly best of the two Mirrors knowledge sources. It thus seems clear that the difference is greatest between SEMANTIC-FEATURES on the one hand, and RELATED-WORDS and WORDS on the other hand.

The comparison in Table (9.6) does not only involve differing models, but also a difference between model settings (the context window settings). A comparison was therefore also conducted between the best SEMANTIC-FEATURES classifiers (column four in Table (9.6)) and its counterpart WORD model, and between the best RELATED-WORDS classifiers (column five in Table (9.6)) and its counterpart WORD model. For the sake of overview with this many experiments, the full tables for the counterpart W comparison have not been included in the chapter. Instead, the overall counts of similar and changed classifications across all target words are given in Table (9.7). The truth values denote correct and wrong classifications, in line with the contingency table of equal and changed cells to be used for a paired statistical test of significance, as described in Chapter (6.4) (p. 134).

Based on the direct comparisons, the counterpart W outperforms the best SF with an OA of .819 (which may be derived from the table by summing the 'True' counts for W, i.e. $888 + 110$, and dividing this on the total, 1219). As opposed to the comparison between the best model settings, however, the direct comparisons do not indicate that the observed decrease from Ws to SFs is statistically significant. As for the best REL-W, its counterpart W outperforms it with three correct instances more (the W overall accuracy is thus $1043/1219$, or .856). But similarly to the main comparison between the best classifiers, this difference is not found to be statistically significant.

| | | **SF** | |
|---|---|---|---|
| | | True | False |
| **W** | True | 888 | 110 |
| | False | 86 | 135 |

| | | **REL-W** | |
|---|---|---|---|
| | | True | False |
| **W** | True | 997 | 66 |
| | False | 63 | 113 |

Table 9.7: A pairwise comparison between each of the two Mirrors-derived models (EXP2 and EXP3) and their WORD counterparts (all target words).

On a general note it may be remarked that the overall accuracies (OAs) would probably be higher if *function words* were included as a source of knowledge, too. Consider for instance *plan*N, being ambiguous (according to the Mirrors method) between *plan*1 ('scheme') and *plan*2 ('level'). The first sense is often recognised with open-class co-occurrences due to word associations such as *legge/forandre planer* 'make/change plans' or *nye planer* 'new plans'. As regards the sense *plan*2, however, the test occurrences of this sense tended to occur in collocational patterns of the type *på (...) plan* 'at the (...) level'. The kinds of content words

within these patterns were quite diverse, however, encapsulated notions spanning from *det mentale plan* 'the mental level', *nasjonalt plan* 'national level', *det praktiske plan* 'the practical level' and *det heroiske plan* 'at the heroic level'. All three knowledge sources had problems with the correct identification of this sense, and it seems quite likely that collocational context features might have recognised this sense more efficiently. The presented experiments, however, focus on the open-class words because we are specifically interested in the context features where we may study the effect of abstracting from lemmas to Mirrors-information about these lemmas.

**The knowledge sources vs. the baseline**

The 'overview' subsection gave a general presentation of which of the three presented knowledge sources perform best. Comparing each of the three models with the baseline (the most frequent sense), Table (9.6) shows that the WORDS classifier and the RELATED-WORDS classifier perform equally well or outperform the baseline with the majority of target words (with eleven of the fifteen target words). By contrast, the SEMANTIC-FEATURES knowledge source only reaches or outperforms the baseline with six of the target words.

**The Mirrors-derived knowledge sources vs. the WORDS model**

Comparing specifically the two Mirrors-based knowledge sources with the best WORDS model for each target word, SEMANTIC-FEATURES is—again—clearly the weakest classifier. Using SEMANTIC-FEATURES in isolation, there is an increase in performance with respect to the WORDS model with four of the target words (*frisk*AJ, *gal*AJ, *lag*N, *rot*N, although the increase is not statistically significant with any of them. With one of the target words there is no change (*fyr*N) and with the remaining ten target words there is a performance drop. Of these, the decrease is found to be statistically significant for *full*AJ, *liv*N, *stemme*N (alpha level=$0.05$).

Comparing the best RELATED-WORDS against the best WORDS model per target word, there is an increase in performance with seven of the target words (*frisk*AJ, *full*AJ, *fyr*N, *gal*AJ, *lag*N, *rot*N, *slag*N). The improvement was statistically significant for *frisk*AJ whereas the remaining increases were, as with SEMANTIC-FEATURES, not statistically significant. As opposed to with SEMANTIC-FEATURES, neither of the cases of a decrease is statistically significant.

**Summing up**

Based on the classification outcomes, there is no statistical evidence to suggest a real difference between the WORD model and the RELATED-WORDS model. This is actually quite encouraging in view of the fact that an information loss should be expected when abstracting to those context lemmas that were automatically sense-tagged. Therefore we will consider this in the next three subsections which address each knowledge source in some detail: were in fact many lemmas 'lost' during training when not all of them are sense-tagged and contribute with Mirrors-information? We will also consider the situation at classification time: is it so that one model knew more context lemmas at classification? Beginning with the best models per knowledge source (cf. Table (9.6) (p. 212)), we will compare their ratio of known lemmas during classification time. As for the two Mirrors-derived knowledge sources (Sections 9.3.4 and 9.3.5) we will also present a direct comparison between the best SF model and its counterpart W model (i.e. when they are trained and tested using exactly the same context window settings) and between the best REL-W model and its counterpart W model. All other things being equal, we will here look at their similarity at classification time by considering how many of the context lemmas were known in both models, unknown in both models or only known in one model. This is of some interest, since we expect *a priori* that there will be an information loss since not all context lemmas are sense-tagged, and in that case we do not know in advance whether any of the models know context lemmas during testing than the other models.

### 9.3.3 EXP1: The WORD classifier

It may be instructive to consider the number of lemmas that contributed to learning in the WORDS model, in order to see the relation between the number of individual lemmas that were registered in the model and the average number of times each lemma actually occurred. The column *average-N* in Table (9.8) (p. 216) shows that with the WORDS model, each unique context feature (i.e. each lemma observed during training) occurs on average in the range of $1.394$ and $2.754$ times in the presented lexical sample. Similarly, the median (the most typical frequency of a context feature) is $1$ for each target word, which echoes the observation from Cucchiarelli and Velardi (2002) (Chapter (3.5.3), p. 56) who argue that most features typically used in WSD are single occurrence phenomena, which, in turn, motivates a search for a more generalised kind of context features.

More specifically, the first column in Table (9.8) shows the TW; the second column shows the model on which the classifier is based. For instance the alphabetically first listed target word in the table, *frisk*AJ, is based on the model $naivebayes(4, 2)$, which means that $4$ context features are collected from the left-

hand side of the target word and 2 words are collected from the right-hand side (this means that maximally 6 lemmas will be collected per target word instance in the training material).

However, the number of unique context features, recorded in the third column in the table show that the 492 collected context features (at token level) are reduced—at type level—to 353 unique context features; in other words, each individual context feature the frequency of occurrence for each of them is not high. This is what is reflected through the mean (the average number of times each context feature occurred) and the median values in columns four and five. These low counts highlight why it might be desirable, in principle, to attempt to generalise from individual context words to classes of semantically related word—that is, if the assumption holds that there will be a gain in such a generalisation and if the Mirrors method is a suitable knowledge resource for this purpose.

Table 9.8: Context features during training and testing: EXP1 (WORDS)

| | | **Training context features** | | | **Test lemmas** | | |
|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Known | Total | Ratio |
| *frisk*AJ | (4-2) | 353 | 1.394 | 1.000 | 49 | 158 | .310 |
| *full*AJ | (2-4) | 1480 | 1.781 | 1.000 | 283 | 779 | .363 |
| *fyr*N | (30-20) | 1385 | 2.043 | 1.000 | 279 | 758 | .368 |
| *gal*AJ | (10-1) | 653 | 1.954 | 1.000 | 120 | 349 | .344 |
| *lag*N | (2-20) | 498 | 1.635 | 1.000 | 77 | 272 | .283 |
| *liv*N | (2-2) | 1604 | 2.308 | 1.000 | 377 | 855 | .441 |
| *plan*N | (2-4) | 436 | 1.500 | 1.000 | 66 | 213 | .310 |
| *rot*N | (2-30) | 1035 | 1.709 | 1.000 | 181 | 499 | .363 |
| *slag*N | (1-10) | 921 | 1.588 | 1.000 | 141 | 457 | .309 |
| *stemme*N | (10-10) | 2422 | 2.754 | 1.000 | 722 | 1343 | .538 |
| *tak*N | (2-2) | 598 | 1.714 | 1.000 | 99 | 313 | .316 |
| *trykke*V | (10-4) | 446 | 1.444 | 1.000 | 59 | 224 | .263 |
| *utsette*V | (30-4) | 1342 | 1.941 | 1.000 | 272 | 674 | .404 |
| *utvalg*N | (30-2) | 725 | 2.025 | 1.000 | 184 | 444 | .414 |
| *valg*N | (4-30) | 1462 | 2.419 | 1.000 | 446 | 805 | .554 |
| | Averaged ratio of known test context features across target words: | | | | | | .376 |
| | | | | | | (standard deviation: | ±0.089) |

So given the model, how many of the lemmas encountered in a test situation were actually known from training? Note that in the present context we specifically speak of the context 'lemma' in a test situation instead of using the more general concept 'context features'. 'Context features' would also be a correct term, but in this table—and in the corresponding tables for each of the two Mirrors-based knowledge sources, which will follow in the two next subsections—we specifically speak of lemmas to underline the point that in a test situation, the classification is based on the lemmas because this is the sort of information that we expect to be able to retrieve from common corpus resources (cf. Chapter (6.5.4), p. 131).

The three last columns per target word in Table (9.8) show the number of known lemmas (at type level) given the total number of context lemmas that emerged during classification. The very last column shows the ratio, whereas the two preceding columns show the corresponding absolute numbers. As can be seen, roughly a third of the encountered lemmas were known to the classifier. The extremely low frequency counts in the WORDS model is what motivates the idea that it might be beneficial to group singular context words together according to relatedness, in order to achieve, in the best case, a more 'compact' learning model.

The results when abstracting to the SEMANTIC-FEATURES and to the RELATED-WORDS of the same training lemmas as in the current subsection will be presented in the following two subsections (Section (9.3.4) and Section (9.3.5)).

### 9.3.4 EXP2: The SEMANTIC-FEATURE classifier

Some details of the SEMANTIC-FEATURE model and the amount of known lemmas at classification time are summed up in Table (9.9) (p. 218); we will begin by explaining the contents of the table. First, an extra column has been added compared to the WORDS Table (9.8), namely the column *Contrib. lemmas* (contributing lemmas). This column shows, for each target word, how many of the lemmas in the training material gave rise to Mirrors-derived SEMANTIC-FEATURES—Seen from the opposite angle, this column indicates how many lemmas were 'lost' because they were not sense-tagged in context. As can be seen from this column, the findings seem to confirm what has also been seen in the automatic sense-tagging experiments (Chapter (7)); approximately half of the lemmas in question were sense-tagged and, thus, contributed with SEMANTIC-FEATURES. (The calculated ratios per target word are not included in the table since they were quite evenly distributed around $50\%$; the maximum ratio is $.612$ and the minimum ratio is $.468$.)

The context features counted in the third column now represent counts of Mirrors-derived SEMANTIC-FEATURES, and not lemmas. Columns four and five show that compared to the best WORD model for each target word, the best SEMANTIC-FEATURE model results in a slight, but consistent, rise in the average frequencies (*average-N*) per context feature. The average is now in the range of $6.489$ and $2.207$. The median is still $1$ with most of the target words; in all we may thus conclude that the median still remains very low. It will be more clear in the controlled experiment in the next chapter whether this very moderate frequency improvement is thus moderate because of the trade-off between added Mirrors information and lost context information when confined to sense-tagged context words.

As regards the classification (the three last columns of Table (9.9)), the ratio of known lemmas is now greater than what was seen with the best WORDS model

Table 9.9: Context features during training and testing: EXP2 (SEMANTIC-FEATURES)

| | | Training context features | | | | Test lemmas | | |
|---|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Contrib. lemmas | Known | Total | Ratio |
| *frisk*AJ | (4-2) | 373 | 2.507 | 1.000 | 216/358 | 94 | 158 | .595 |
| *full*AJ | (2-1) | 639 | 2.900 | 1.000 | 429/801 | 188 | 396 | .475 |
| *fyr*N | (4-2) | 247 | 2.421 | 1.000 | 128/248 | 70 | 121 | .579 |
| *gal*AJ | (1-2) | 225 | 3.538 | 1.000 | 106/175 | 39 | 86 | .453 |
| *lag*N | (100-75) | 1469 | 6.489 | 2.000 | 1133/2421 | 673 | 1453 | .463 |
| *liv*N | (2-2) | 1286 | 5.607 | 1.000 | 928/1604 | 483 | 855 | .565 |
| *plan*N | (2-2) | 328 | 2.351 | 1.000 | 177/306 | 83 | 147 | .565 |
| *rot*N | (1-4) | 198 | 2.207 | 1.000 | 115/204 | 57 | 104 | .548 |
| *slag*N | (1-4) | 442 | 2.722 | 1.000 | 256/468 | 115 | 236 | .487 |
| *stemme*N | (4-1) | 728 | 5.312 | 1.000 | 456/845 | 234 | 428 | .547 |
| *tak*N | (1-2) | 386 | 3.207 | 1.000 | 240/467 | 115 | 244 | .471 |
| *trykke*V | (50-75) | 1636 | 6.417 | 1.000 | 1275/2390 | 654 | 1257 | .520 |
| *utsette*V | (4-2) | 285 | 2.386 | 1.000 | 157/297 | 73 | 138 | .529 |
| *utvalg*N | (20-1) | 425 | 3.318 | 1.000 | 255/531 | 167 | 307 | .544 |
| *valg*N | (10-2) | 580 | 3.317 | 1.000 | 354/680 | 201 | 377 | .533 |
| | | | | | Averaged ratio of known test context features across target words: | | | .525 |
| | | | | | | (standard deviation: | | $\pm$0.045) |

(Table (9.8)). Bear in mind that the best WORDS model may have a different context window setting than the best SEMANTIC-FEATURES model, therefore it is not methodologically perfectly accurate to compare the known lemmas in the two models (although the ratio is a good approximation).

In order, therefore, to achieve a *direct* comparison of the loss or gain at classification time in abstracting to SFs, its WORD counterpart may be considered. Since the best SF model and its counterpart WORD model use the same context window setting, they will encounter exactly the same context lemmas in a test situation. Figure (9.2) uses a 100% stacked bar chart to depict the relationship between known and unknown test lemmas in the two models. Each of the fifteen bars represents the full set of test lemmas (at type level) encountered in the classification of a target word. Test lemmas known in both models are denoted as $TT$ ('true-true') in the figure, lemmas not known in any of the models are denoted $FF$ ('false-false'), lemmas only known in the W model are denoted $TF$ ('true-false') and lemmas only known in the SF model are denoted $FT$ ('false-true').

Figure (9.2) suggests that on the whole, the best SF model and its counterpart W have shared knowledge or a lack of such: between $90\%$ and $70\%$ of the context lemmas in a test situation are known ($TT$) or unknown ($FF$) in both models. The most interesting ratio for us would concern the two areas where the models differ, that is, the $FT$ area and the $TF$ area (the two topmost areas). These two areas are too small to allow for generalisation; thus the main conclusion suggested by this figure is rather that although the OA difference between the two models is statistically significant, this difference does not seem to find its explanation in a tremendeous difference between what the two models knew (they majority of test

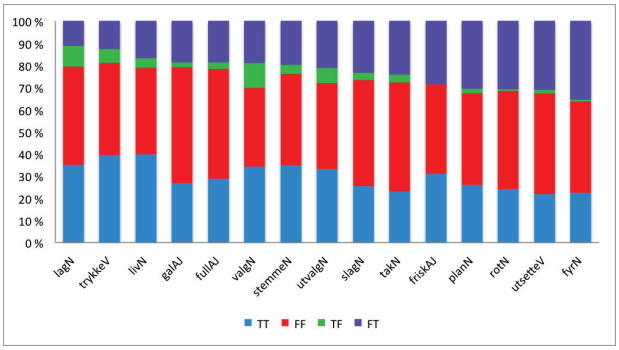



Figure 9.2: Known and unknown lemmas at classification time in the best SF model and its counterpart W model, as a 100% stacked bar chart.

lemmas were known or unknown in *both* models). It is also interestingly to note that the amount of test lemmas only recognised by the SF model ($FT$s) is greater than the opposite group (lemmas only known in the WORD model). Although the material is too small to give much emphasis to this, it would suggest that the words that are lost during training because they were not sense-tagged do not necessarily have a great impact on the ability of the SFs to recognise words in a test situation.

An analysis of the $TF$s (the lemmas only known in the W model), incidentally, indicate that across all fifteen target words more than half of the lemmas were nouns. In other words, the main problem for automated word-alignment, and thus the main part of speech where we lose input to the Mirrors method, is nouns. The $TF$ lemmas tended to be hapax words and we often find proper names (*Jahve*N), compounds (*innflytelsesrik*AJ 'influence-rich', i.e. 'influential'), *suppeskilpadde*N 'soup turtle') and a variety of nouns that are akin to a kind of function words, such as *nr.*N ('no.; number') or *a*N (which was probably part of an alphabetised list in a text).

A manual analysis of the classification outcomes suggest thes same tendency that was exemplified in the discussion of model selection experiments (Section (9.2.2)): often the frequencies are not much higher per context feature than with WORDS, and some of the SFs seem to be too general to pull unequivocally in the same direction as what we find in a WORD model. Although it has not been quantified, it especially seems that verb SFs unite too many word senses, which is also natural since it was known in advance that verb SFs often have quite large denotations (cf. the discussion of SFs and REL-Ws in Chapter (6)).

Moreover, it can be remarked that an underlying expectation of this thesis would be that if the Mirrors-based knowledge sources are plausible, and if they

| | | SF | |
|---|---|---|---|
| | | True | False |
| W | True | 898 | 140* |
| | False | 76 | 105 |

Figure 9.3: A pairwise comparison between the best W models of EXP1 and the best SF models of EXP2. A starred cell count means that the decline/increase is statistically significant ($\alpha$ = .05).

really do add more information than what was already directly present in the text corpus, then it is natural to expect the SEMANTIC-FEATURE-based model to be able to classify more test instances than a traditional WORD model. A count among the 1219 total test instances revealed, however, the opposite, although the following difference was not found to be statistically significant. The WORD-based model answered *don't know* with 23 test instances whereas the SEMANTIC-FEATURE model answered the same 31 times (cf. Table (9.10)). Since this observed difference is not statistically significant with the conventional $\alpha \leq 0.05$, we will not pursue this further in any detail; we will instead conclude that the present material shows no evidence to suggest that SEMANTIC-FEATURES could have a markedly clear potential to 'fill the gaps' where the traditional WORD classifier does have any knowledge at all. It will be pointed our at the same time that as the same applies to RELATED-WORDS, we will not pursue this any further with RELATED-WORDS, either (regarding the classification choices of the WORDS classifier and the RELATED-WORDS classifier, the WORD based model answered *don't know* with 23 test instances whereas the RELATED-WORDS model answered the same 27 times).

Table 9.10: EXP1 (W) vs. EXP2 (SF): classification answers

| | W | SF |
|---|---|---|
| Classifications made | 1196 | 1188 |
| *Don't know* | 23 | 31 |
| Total instances | 1219 | 1219 |

We will conclude this subsection by showing some illustrative examples of the behaviour of the SEMANTIC-FEATURE classifier. The target word instance in Example (14) could not be sense-tagged with the WORD model because no context lemmas were known whereas it was correctly classified by the SEMANTIC-FEATURES classifier. The example illustrates both a strength and a weakness of attempting to obtain a generalisation using SFs. On the one hand, the two recognised context lemmas can be traced back to quite satisfactory semantic relations through the Mirrors method. On the other hand, the frequencies that will be shown for the

relevant SFs illustrate the general observation that the frequencies are lower than what was perhaps expected even though co-occurrences are 'counted toghether' through SFs.

(14) Renna som fører regnvann fra ***taket7*** til Eleonorah og ned i et digert kar i regntida. (TB1)
The drainpipe that carries the water from Eleonorah's **roof** and down into a large barrel in the rainy season.

The WORDS classifier based itself on the two nearest open-class lemmas on each side, the noun *regnvann* 'rain water', the verb *føre* 'carry', the proper name *Eleonorah* and the adjective *diger* 'large'. By contrast, the SEMANTIC-FEATURES model (using one lemma less on the left-hand side) rested on the same lemmas except the name *Eleonorah*.

Of these, the noun *regnvann* 'rain water' and the adjective *diger* 'large' were indirectly known from training, not because any of them actually occurred but because they share SFs with actually occurring lemmas. The noun was recognised in virtue of sharing a SF with the actually co-occurring context lemma *regn*N 'rain'. Thus, this example illustrates that the use of Mirrors-derived related word senses (in this case through SEMANTIC-FEATURES) may lead to the inclusion of plausible, potential co-occurrences that were not actually attested in the training corpus material.

The adjective was recognised through five SEMANTIC-FEATURES. Their observed frequencies per target word sense are listed in Table (9.11) two illustrate two tendencies found in the SF material: A theoretical point of abstracting from Ws to SFs was to obtain higher counts for more reliable statistics, but the observed frequencies are often actually on the low side. Second, when SFs did generalise during training (i.e. two or more lemmas were united in virtue of sharing a SF), the resulting counts more often seem to pull in different directions with regard to the target word sense. The $[large1|liten1]$ in the table, for instance, is found to occur with all three target word senses, and it thus does not have a strong 'pulling strength' when the Naive Bayes formula is applied to compute the probability of each sense given the identified, known context evidence.

So to the extent that SFs 'blur' the contextual difference between target word senses (since the W model outperforms the REL-W), should this be attributed to the plausibility of the Mirrors method? It seems that the main problem in using the Mirrors method concerns the available input: the amounts of translational data are on the low side and it seems that some word alignment errors create many problems in the Mirrors method. Another interesting observation is that it seems sometimes problematic to assume that evene linguistically well-motivated

Table 9.11: The training frequencies of five adjective SEMANTIC-FEATURES that created a link to a previously unseen test context word, *diger* 'large'

| SF | *tak*2 | *tak*4 | *tak*7 | Contributing training lemmas |
|---|---|---|---|---|
| $[enormous1\|mektig1]$ | 0 | 0 | 1 | {*enorm*AJ } |
| $[enormous1\|veldig1]$ | 0 | 0 | 1 | {*enorm*AJ} |
| $[great1\|stor1]$ | 1 | 2 | 2 | {*enorm*AJ *lang*AJ *mange*AJ *høy*AJ } |
| $[large1\|liten1]$ | 4 | 2 | 3 | {*lang*AJ *øvre*AJ *mange*AJ *gammel*AJ *liten*AJ} |
| $[huge1\|hoey1]$ | 0 | 2 | 1 | {*enorm*AJ *høy*AJ } |

relations in the paradigmatic dimension have similar distributional properties in the syntagmatic dimension.

A particularly good example of this is found with one of the test instances of the target word *frisk*AJ (ambiguous between the senses of 'fresh' (*frisk*1) and 'healthy' (*frisk*4)). The lemma *bli*V is intuitively characteristic of the 'healthy' sense, *frisk*4, as in the collocation ***bli*** *frisk igjen* 'get well soon'. When comparing the best SEMANTIC-FEATURES model of *frisk*AJ against the counterpart WORD model, the co-occurrence *bli*V was seen in the WORD model twice with the sense *frisk*4 and zero times with *frisk*1. Although low-frequent, this observation could thus be said to be in line the simple intuition that this lemma is characteristic of the 'healthy' sense, especially given the very narrow context window that resulted from the model selection phase. Considering the classification output with the counterpart WORD model, this lemma is also among the ten most informative co-occurrences based on how strongly it 'pulls' in one direction (the 'pulling strength' is computed as the ratio between its highest probability value for a target word sense and the sum of its probabilities per target word sense).

But since the verb *bli*V ('become, get') has a quite general meaning, it is not necessarily beneficial to tentatively include words that are related to it. Considering the classification output with the SEMANTIC-FEATURES model, the semantic feature $[get1|bli1]$ is in fact the lowest-ranked context feature in the entire model in terms of its contribution. The reason is that this SEMANTIC-FEATURE, being quite general, was found to occur nine times with *frisk*1 and four times with *frisk*4; its MLE values then yield a ratio of only $0.51$. In other words, one of the ten most informative lemmas in the WORD model gives rise to the least informative SEMANTIC-FEATURE of all recorded features in the SF model.

### 9.3.5 EXP3: The RELATED-WORDS classifier

Considering, now, the classification results based on RELATED-WORDS, our starting point from the Overall Accuracies (OAs) in Table (9.6) (p. 212) is that there is not statistically significant difference between training on traditional WORDS as

| | | REL-W | |
|---|---|---|---|
| | | True | False |
| W | True | 953 | 85 |
| | False | 87 | 94 |

Table 9.12: A pairwise comparison between the best W models of EXP1 and the best REL-W models of EXP3. A starred cell count means that the decline/increase is statistically significant ($\alpha$ = .05).

opposed to on the RELATED-WORDS, even when we know that some context lemmas are lost because they are not sense-tagged and cannot contribute in the RELATED-WORDS model. Recall from Table (9.7) (p. 213) that there was also no statistically significant difference if the best RELATED-WORDS model per target word is compared directly against its counterpart WORD model.

Details of the training and known test lemmas with the best RELATED-WORDS model per target word is summed up in Table (9.13) (p. 223). As with SEMANTIC-FEATURES, there are four columns related to training, because we also consider the column of *Contrib. lemmas* (contributing lemmas). These results concord with the findings from the SF model; approximately half of the lemmas were sense-tagged and could contribute with RELATED-WORDS (the calculated ratios for each target word are not included; the maximum ratio is .606 and the minimum ratio is .495).

Table 9.13: Context features during training and testing: EXP3 (RELATED-WORDS)

| | | Training context features | | | | Test lemmas | | |
|---|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Contrib. lemmas | Known | Total | Ratio |
| *frisk*AJ | (2-30, 15) | 2203 | 2.554 | 1.000 | 769/1376 | 298 | 700 | .426 |
| *full*AJ | (4-1, 40) | 1901 | 2.327 | 1.000 | 655/1250 | 252 | 642 | .393 |
| *fyr*N | (30-10, 10) | 1836 | 2.535 | 1.000 | 606/1193 | 247 | 623 | .396 |
| *gal*AJ | (1-2, 50) | 403 | 1.514 | 1.000 | 106/175 | 24 | 86 | .279 |
| *lag*N | (30-10, 4) | 1348 | 2.142 | 1.000 | 412/832 | 169 | 455 | .371 |
| *liv*N | (1-2, 50) | 2024 | 3.175 | 2.000 | 741/1265 | 341 | 670 | .509 |
| *plan*N | (10-2, 20) | 1448 | 1.820 | 1.000 | 444/812 | 166 | 398 | .417 |
| *rot*N | (4-30, 20) | 1742 | 2.266 | 1.000 | 599/1092 | 225 | 524 | .429 |
| *slag*N | (30-4, 40) | 2916 | 3.263 | 1.000 | 1193/2192 | 499 | 1142 | .437 |
| *stemme*N | (10-10, 10) | 3032 | 5.286 | 2.000 | 1274/2422 | 688 | 1343 | .512 |
| *tak*N | (2-2, 10) | 1077 | 1.780 | 1.000 | 302/598 | 118 | 313 | .377 |
| *trykke*V | (20-30, 15) | 1995 | 2.515 | 1.000 | 690/1243 | 283 | 655 | .432 |
| *utsette*V | (10-20, 30) | 1839 | 2.538 | 1.000 | 630/1202 | 253 | 635 | .398 |
| *utvalg*N | (10-30, 15) | 1351 | 2.811 | 1.000 | 420/824 | 213 | 549 | .388 |
| *valg*N | (20-10, 30) | 2110 | 2.799 | 2.000 | 738/1382 | 366 | 743 | .493 |
| | | | | | Averaged ratio of known test context features across target words: | | | .417 |
| | | | | | (standard deviation: | | | $\pm$0.059) |

The context features (in the third column) now consist of Mirrors-derived RELATED-WORDS and not lemmas. Columns four and five show that compared

to the best WORD model, the best RELATED-WORD models result in a slight rise in the average frequencies (*average-N*) per context feature. Considering the average per target word, the averages are in the range of $5.286$ and $1.514$. Compared to the median in the WORD model presented in EXP1, the median is now $2$ with a fifth of the target words. We thus see a slight increase, but overall the median remains extremely low. The absence of a convincing increase in frequencies may be the explanation for why there seems to be no effect of generalising to RELATED-WORDS. In that case, it will be more clear in the controlled experiment in the next chapter whether this very moderate frequency improvement is thus moderate because of the trade-off between added Mirrors information and lost context information when confined to sense-tagged context words, therefore this will not be further pursued in the current chapter.

Table (9.13) indicates that the average amount of known lemmas at classification time was slightly higher with RELATED-WORDS than with WORDS when considering the ratios (bear in mind that the absolute numbers are not directly comparable, since the models being compared may have different context windows). The average ratio across target words increases from $.376$ to $.417$. By contrast, when considering the W counterpart of the best REL-W models, the ratio of known lemmas during classification is $.417$ with REL-Ws and $.436$ with Ws; in other words there has been a slight loss of known lemmas when generalising from Ws to the REL-Ws of those context words that were sense-tagged. At the same time the best REL-W model learns roughly twice as many context features as in its counterpart W model, based on roughly half the amount of context lemmas.

We will consider one target word, *frisk*AJ, to illustrate what is meant. Column six in Table (9.13) shows that on the basis of the $1376$ actually occurring context lemmas, 'only' $769$ contributed with Mirrors-derived information; so roughly half of the lemmas at type level were lost. Column three shows that with half the amount of actually co-occurring context lemmas, the best RELATED-WORD model obtained $2203$ context features (word senses). By contrast, the WORD counterpart model had $1376$ unique context features in its model (we do not include the full tables for counterpart models in the chapter). This means that as far as learning is concerned, the 'gain' was actually tremendous for the RELATED-WORD model. But as far as the classification rates are concerned, the comparison between the same RELATED-WORDS model and its counterpart WORD model indicates—put bluntly—that there seems to have been no real effect of learning these added context features, since the WORD classifier knew more of the test lemmas than the RELATED-WORDS model and since the classification accuracy of the two models indicate no statistically significant difference.

Figure (9.4) compares whether context lemmas were known or unknown in the best RELATED-WORDS model and in its counterpart WORD model, using a 100% stacked bar chart. As in Figure (9.2) (p. 219), the context lemmas that were known

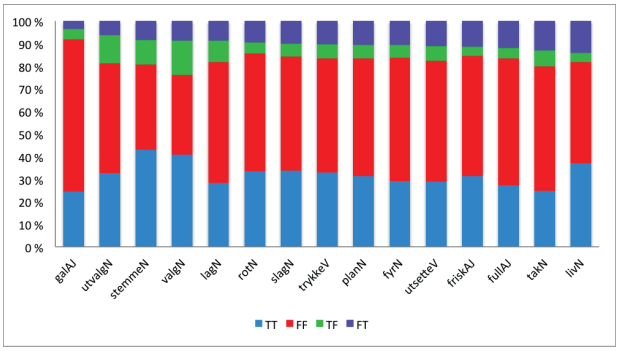



Figure 9.4: Known and unknown lemmas at classification time in the best REL-W model and its counterpart W model, as a 100% stacked bar chart.

in both models are denoted as $TT$, ('true-true') in the figure, lemmas not known in any of the models are denoted $FF$ ('false-false'), lemmas only known in the W model are denoted $TF$ ('true-false') and lemmas only known in the SF model are denoted $FT$ ('false-true').

Judging from the figure, most of the test lemmas are either known in both ($TT$) or unknown in both ($FF$); in other words the statistical finding that these two models are not statistically different in terms of overall accuracies seems to find its explanation in the fact that there is also no marked difference in the context lemmas that they have learnt. For the sake of WSD and the hope to improve the results by adding paradigmatic knowledge, such a finding is not especially promising, since this implies that the two knowledge sources do not appear to possess complementary knowledge, at least not to a statistically significant extent. But as far as the plausibility of the Mirrors method is concerned, this finding is quite encouraging, since this means that even when some context words are lost, the RELATED-WORDS knowledge source still seems to obtain the same, necessary information as a traditional WORDS model.

Summing up, it would seem that the generalisation from WORDS to RELATED-WORDS has, as such, a tremendous effect in that with only half of the word co-occurrences the RELATED-WORDS classifier still learns twice as many potentially relevant context features as its counterpart W model. However, the classification outcome suggests that at classification time, the RELATED-WORDS model and its corresponding WORD model seem to have learnt very much the same, since there was only a marginal increase in the ratio of known lemmas and since the amount of lemmas that are known or unknown in *both* models dominates.

As with the SEMANTIC-FEATURE, we will conclude this subsection with an

illustration of the behaviour of the RELATED-WORDS model through some concrete classification outcomes. Example (15) illustrates a test instance for the target word *gal*AJ, ambiguous between a sense of 'crazy' (*gal*1) and 'wrong' (*gal*2)[4]. The example, which pertains to a judicial sentence, was correctly classified by the WORD classifier as *gal*2 whereas the RELATED-WORD classifier answered *Don't know*.

(15) En ***gal2*** dom — uansett i hvilken retning — er åpenbart meget uheldig her. (LSPL1)
A **wrong** sentence—regardless of direction—is obviously very unfortunate here.

With this target word the best WORD model was $naivebayes(10, 1)$ and the best RELATED-WORDS model was $naivebayes(1, 2)$. The RELATED-WORDS classifier thus collects the nearest left-hand side lemma and the two nearest right-hand side open-class lemmas. The lemmas that are registered in the Mirrors lexicon are converted into their possible Mirrors senses; for instance, the context lemma *retning*N ('direction') exists in the Mirrors word bases and has no less than seven senses; *retning*1—*retning*7. Each such sense is looked up in the RELATED-WORDS classification model, since each entry in this model is a word sense. Of the eleven lemmas retrieved from the context of the test instance with the WORD model, four lemmas were recognised: *sterk* 'strong', *skulle* 'should', *bli* 'become' and *god* 'good'. By contrast, the best RELATED-WORDS model had three lemmas at its disposal (*mulig* 'possible', *dom* 'sentence' and *retning* 'direction'), of which neither were known from training.

We will also consider an example of a target word instance which was not recognised with the traditional WORD model (a $naivebayes(2, 4)$ model) whereas the best RELATED-WORDS model ($naivebayes(10, 2)$) did classify it correctly. The target word is *plan*N, being ambiguous in the Mirrors method between the sense of 'scheme' (*plan*1) and 'level' (*plan*2). The correct sense is *plan*1.

(16) Etter at komiteen hadde fått seg forelagt kjensgjerningene ved den brasilianske ***planen*1** og ved andre kontroversielle prosjekter i Indonesia, India og (. . .). (LT1T)
After hearing evidence on the Brazilian **scheme** and other controversial projects in Indonesia, India and (. . .).

Whereas the WORDS classifier recognised none of the six context lemmas at its disposal, the RELATED-WORDS classifier recognised five of the twelve lemmas it

[4]The complete sentence context, taken from the ENPC corpus online to get the complete sentence translation, had an extremely free translation in the corpus. For the non-Norwegian reader the example has therefore been given a word by word translation by the author instead.

had at its disposal. The lemmas known in the RELATED-WORDS model were *prosjekt* 'project', *utvikling* 'development', *la* 'let', *ha* 'have' and *få* 'get'. The noun *prosjekt*N was recognised because its word sense *prosjekt*1 was recorded three times during learning as a co-occurrence of *plan*1 (and it never co-occurred with *plan*2). Of the known lemmas (or rather, the senses of the five lemmas above) the word sense *prosjekt*1 contributed the most to the correct classification of this instance. It may be noted that *project*N never occurred in the actual context of the target word during learning (which is why the WORDS model was unable to recognise it). The word sense *prosjekt*1, however, was recorded thrice during training; all three times because it is a related word to the word sense *plan*1 itself (which means that this target word sense sometimes co-occurred with itself). Using the RELATED-WORDS definition in Chapter (6.3.4) (p. 119), the sense *plan*1 gives rise to a set of related word senses consisting of {*plan*1, *prosjekt*1}.

### 9.3.6 Combining classifiers

Since the analysis of the three individual classifiers has revealed that they do make different classification decisions (although not necessarily statistically significant different decisions), the classifiers were finally combined in a voting scheme in which the most confident classifier gets to classify the test instance in question. 'Confidence', in this context, refers to the probability value of the most probable target sense, as produced with the Naive Bayes formula (Chapter (6)). Thus, the classifier with the highest $p$-value for a target word sense with respect to a particular test instance is allowed to classify this instance.

Obviously, classifiers may be combined using different strategies—Pedersen (2000) explored majority voting (choosing the sense for which most classifiers voted) whereas Hoste, Hendrickx, Daelemans and Van Den Bosch (2002) used weighted voting (see e.g. Navigli, 2009, p. 22, for a more detailed description of combinations). The choice to use confidence as a metric was motivated, above all, by a need to make the combination as simple and transparent as possible. For instance, by keeping the classifiers as separate models their frequencies are also kept apart, which means that we avoid any potential confusion as to whether the combined classifiers interact in complex ways. It was considered to combine the classifiers with majority voting, but in that case one would need an alternative solution should it happen that more than two target word senses are possible and where none of the senses get the majority vote (recall that we only have three classifiers to combine). Another option was to follow Hoste, Hendrickx, Daelemans and Van Den Bosch (2002) in giving extra weight to the classifier with the highest overall accuracy from the model selection experiments; but in that case, too, one might need a solution in cases where more than one classifier was ranked with the best overall accuracy.

Thus, the choice therefore made to let the most *confident* classifier choose the sense for a given test instance. Letting the most confident classifier classify has the bonus effect that it highlights whether there is any statistical gain in adding generalised information from the Mirrors method. In the ideal case—if the Mirrors method is plausible and if a Mirrors-based classifier is enriched compared to a traditional WORD model—it would be natural to expect that the two Mirrors-based classifiers should be more confident than the WORD model, and thus get to classify more instances. Thus, the combined classifier should allow us to identify whether one knowledge source leads to more confident classifiers (by counting which classifier classifies the most instances), and whether the most confident is also the most correct.

Table 9.14: Overall Accuracy for knowledge sources individually and combined (#=absolute counts, %=relative proportion

| TW | MFS | WORDS | | SEMANTIC-FEATURES | | RELATED-WORDS | | Comb. | |
|---|---|---|---|---|---|---|---|---|---|
| | | # | % | # | % | # | % | # | % |
| *frisk*AJ | 50.0 | 20/36 | .556 | 22/36 | .611 | 29/36 | .806 | 22/36 | .611 |
| *full*AJ | 94.7 | 156/189 | .825 | 140/189 | .741 | 165/189 | .873 | 166/189 | .878 |
| *fyr*N | 68.0 | 18/25 | .720 | 18/25 | .720 | 19/25 | .760 | 18/25 | .720 |
| *gal*AJ | 70.6 | 33/51 | .647 | 34/51 | .667 | 34/51 | .667 | 37/51 | .725 |
| *lag*N | 64.7 | 12/17 | .706 | 13/17 | .765 | 13/17 | .765 | 14/17 | .824 |
| *liv*N | 98.2 | 383/398 | .962 | 372/398 | .935 | 376/398 | .945 | 387/398 | .972 |
| *plan*N | 74.5 | 38/47 | .809 | 32/47 | .681 | 37/47 | .787 | 39/47 | .830 |
| *rot*N | 80.0 | 16/25 | .640 | 19/25 | .760 | 18/25 | .720 | 16/25 | .640 |
| *slag*N | 54.4 | 34/57 | .596 | 29/57 | .509 | 35/57 | .614 | 30/57 | .526 |
| *stemme*N | 94.4 | 144/144 | 1.000 | 135/144 | .937 | 142/144 | .986 | 139/144 | .965 |
| *tak*N | 50.5 | 83/111 | .748 | 74/111 | .667 | 78/111 | .703 | 82/111 | .739 |
| *trykke*V | 71.4 | 16/21 | .762 | 11/21 | .524 | 15/21 | .714 | 12/21 | .571 |
| *utsette*V | 69.700005 | 27/33 | .818 | 22/33 | .667 | 26/33 | .788 | 23/33 | .697 |
| *utvalg*N | 55.0 | 20/20 | 1.000 | 17/20 | .850 | 20/20 | 1.000 | 20/20 | 1.000 |
| *valg*N | 57.8 | 38/45 | .844 | 36/45 | .800 | 33/45 | .733 | 36/45 | .800 |

Table (9.14) includes the same three columns that are given in Table (9.6) (p. 212) and adds as a fourth column the result when combining the classifiers. Comparing each of the individual classifiers with the combined classifier, the combined classifier was the overall best classifier with the highest accuracy (or with the same value as the winning accuracy) with nine of the fifteen target words. As with the WORD-based classifier and the classifier based on RELATED-WORDS, the combinatorial classifier performed equally well or outperformed the baseline with the majority of target words (with ten of the fifteen target words).

In order to assess the contribution of each individual knowledge source in the combined classifier, a simple count was made to find how many instances each knowledge source got to classify (which of them was most often the most confident); and of these, how many were correct? Table (9.15) sums up the contribution

of each individual classifier in the combined classifier by presenting the sum of the classifications made by each knowledge source across all fifteen target words. The column entitled *Classifications made* in Table (9.15) shows the sum of classifications made by each knowledge source in the combinatorial classifier and the total sum of classifications to be made across the target words; after the absolute values the ratio is given.

Table 9.15: The contribution of each individual classifier in the combined classifier

| Knowledge source | Classifications made | | Correct classifications | |
|---|---|---|---|---|
| WORD | 466/1219 | .382 | 419/466 | .899 |
| SEMANTIC-FEATURES | 328/1219 | .269 | 258/328 | .787 |
| RELATED-WORDS | 425/1219 | .349 | 390/425 | .918 |

As column two shows, none of the classifiers account for a *prima facie* markedly greater portion of the instances to be classified. Regarding the number of *correct classifications* given the number of classifications made (precision) in column three, the RELATED-WORDS classifier emerges as the most precise classifier, with 390 correct classifications out of 425 classifications made, which yields a precision of .918. By comparison, the WORDS classifier has a marginally lower precision level, .899, whereas the SEMANTIC-FEATURES classifier has a level of .787.

A plot of the overall accuracy (OA) from Table (9.15) sorted by increasing OA for WORDS is depicted in Figure (9.5). The figure shows that for each word, there are no extreme OA differences between the models although the SF (red circles) accuracies are generally found below the other models (recall that for three of the target words, this decrease was also found to be statistically significant).

Figure (9.6) depicts a plot of the OA from Table (9.15) sorted independently for each model (including the combinatorial classifier) in order to show the span between accuracies within the same model. It appears that the distribution of OA among the words is fairly similar for all the models, with SFs performing worse than the rest.

On the whole, the combinatorial classifier seems to yield the same outcome as seen with the individual classifiers; there seems to be no significant difference, whether in terms of a loss or a gain, between WORDS-based classification and RELATED-WORDS-based classification.

## 9.4 Discussion and Conclusion

It must be borne in mind that all the presented classifiers use open-class words only. The classification analysis has shown that for the sake of WSD perform-

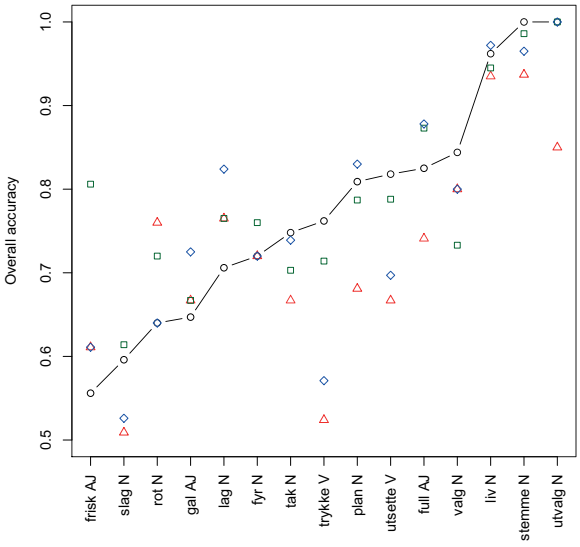



Figure 9.5: Overall Accuracy from Table (9.15) sorted by W. Legend: black=W, red=SF, green=REL-W, blue=combination.

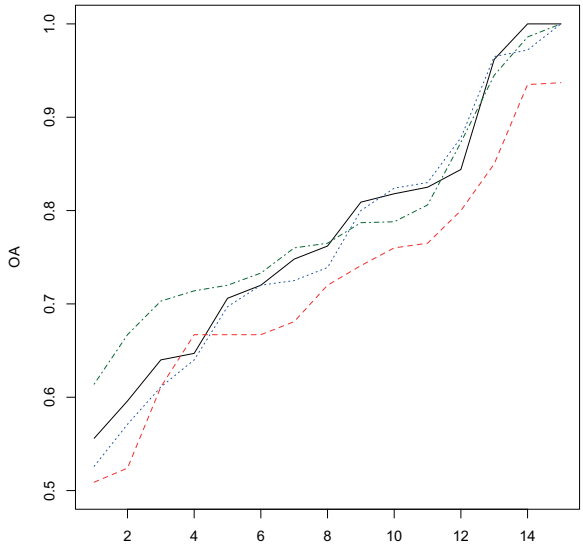



Figure 9.6: Overall Accuracy from Table (9.15) sorted by each model independently. Legend: black=W, red=SF, green=REL-W, blue=combination.

ance in itself (and not for the sake of evaluating the Mirrors method) the overall accuracies would probably have been improved if local collocations had been included as a source of knowledge.

The presented analysis was based on the $n$ nearest WORDS, of which only a subset was sense-tagged and, hence, provided extra knowledge from the Mirrors method. Applying the three knowledge sources in isolation, the results indicate that there is no statistically significant difference between a 'traditional' WORD-based classifier and one that is based on RELATED-WORDS, whether in terms of a loss or a gain. Furthermore, regarding each knowledge source with its best context window settings, the SF-based classifier performs significantly worse that the traditional WORDS-based classifier as well as the other Mirrors-based RELATED-WORDS model.

The analysis suggests that the frequency counts for a context feature may be slightly higher in the SEMANTIC-FEATURE model, but they sometimes pull in a contradictory direction compared to a simple WORDS model. Moreover, it has been shown that the contribution of individual context features does not in itself depend on higher frequencies but on the ratio between the MLEs: the higher the ratio, the more it 'pulls' in the direction of a given sense of the target word. Since this applies even with low-frequent context features, the most important factor seems to concern a notion of precision: the most important issue is that context features must pull in the same direction.

The analysis of the RELATED-WORDS model shows that this model learns more word senses during training than a WORD model learns context lemmas (under equal context window settings), although the REL-W model bases loses half of the context windows during learning (since not all context lemmas are sense-tagged). This implies that a certain generalisation is indeed present during training. But at classification time, the two models (the REL-W and its counterpart W) predominantly know the context lemmas and their Overall Accuracies (OAs) are also without statistically significant differences.

So the findings of this chapter support previously seen evidence, both in (Chapter (7)) and in previous work: approximately half of the lemmas in the ENPC are sense-tagged automatically, which means that quite a lot of information is in principle lost when focussing on Mirrors-specific context information. Seen from this point of view, it is actually quite encouraging that there is no greater loss in the classification accuracy when abstracting from the nearest words to Mirrors-derived information.

So for the sake of WSD and the hope to improve the results by adding paradigmatic knowledge, the missing *gain* may appear disappointing, But as far as the plausibility of the Mirrors method is concerned, this finding is quite promising, since this means that even when some context words are lost, the RELATED-WORDS knowledge source still seems to obtain the same, necessary information

as a traditional WORDS model.

As for the contribution of the different knowledge sources in a combination scheme where the most confident gets to classify (taking its p-value as an indicator of how well-informed the classifier was about the context of the test instance in question), it is not easy to judge whether it would be legitimate to expect a much higher contribution ratio on the part of the two two Mirrors-based classifiers. As was discussed in Chapter (3.5.2), Pedersen (2002) compared several WSD systems that participated in SENSEVAL-2. Pedersen found that if several systems are largely in agreement, then there is little benefit in combining them since they will simply reinforce each other (*combinatorial benefits* will, on the other hand, improve performance significantly if combined appropriately).

On the whole, the findings of the present chapter motivate a controlled study in the next chapter in order to measure whether the moderate improvement seen in this chapter hinges on the trade-off between added Mirrors information and lost context information.

# CHAPTER 10

# THE DIRECT LOSS OR GAIN IN ADDING INFORMATION FROM THE MIRRORS METHOD

## 10.1 Chapter introduction

The previous chapter took a traditional WORD classifier as its starting point, testing the effect of replacing the actually occurring lemmas with information from the two Mirrors-derived knowledge sources—SEMANTIC-FEATURES and RELATED-WORDS, when available. The results showed that since not all context lemmas are sense-tagged, some context information is simply lost when moving from context WORDS to Mirrors-derived knowledge found among the same set of context words. Under such circumstances it is hard to assert whether any theoretical gain in adding Mirrors-derived information may have been countered by the information loss caused by the fact that some words that were observed in the WORD model were missing from the corresponding Mirrors-derived models.

This chapter therefore presents a purely theoretical evaluation which singles out the $n$ nearest *sense-tagged* context words that are sense-tagged. The purpose is to remove the 'loss of information' factor and thus to isolate the theoretical effect of replacing the actually occurring words with Mirrors-derived information about the same words, in order to be able to say something about the quality, or plausibility, of the Mirrors-derived information (EXP5, EXP6, EXP7). We will also present a controlled experiment aimed at testing the plausibility of the Mirrors word senses (EXP8). In this experiment the sense partitions are ignored, i.e. Mirrors-derived semantic information from all the possible senses of a context word are considered

(explained in more detail in Chapter (6.4.2)).

As specified in Chapter (6.4.2), the context information to be applied in the four experiments of the present chapter then contains the following:

**EXP5** WS: The [$\pm$ $n$] nearest WORDS that were sense-tagged.

**EXP6** SFS: The SEMANTIC-FEATURES (SFS) derived from all the actually occurring context lemmas in EXP5.

**EXP7** REL-WS: The word sense associated to each context word in EXP5 together with the RELATED-WORDS of each such context word sense.

**EXP8** UNION REL-WS: The union of *possible* Mirrors word senses (irrespective of which is predicted according to the automatic sense-tagging) for each context word in EXP5 together with the RELATED-WORDS associated to each such word sense.

The last experiment (EXP8) will be performed on the basis of the same context window settings as in EXP7, as it would take this chapter a bit too far to develop own classifiers for this experiment using model selection. Moreover, it appears most interesting to compare the UNION classifier in EXP8 against the RELATED-WORDS classifier in EXP7, since the latter may said to contain 'sense-specific' RELATED-WORDS whereas the former contains *not* sense-specific RELATED-WORDS associated to a context lemma in any possible sense.

The experimental framework in this chapter means that every context lemma in the WORD model is also present in the corresponding Mirrors-derived models, but in a semantically analysed form. For instance, then, the stacked bar charts in the last chapter (Figures 9.2, p. 219 and 9.4, p. 219) should now *not* contain any $TF$ lemmas. If something is known in the WORD model the same should also be known in the Mirrors-derived model under equal context window settings (but maybe with different frequency counts). Recall the English example given in Figure (6.9) (p. 123), Chapter (6), which used a context example for the English target word *bill*. One of the context words found in that example sentence was *telephone*N, and this context word was also automatically sense-tagged as *telephone*2. So in the WORD model, the semantically unanalysed lemma *telephone*N is recorded as a context word of *bill*N. In the corresponding SEMANTIC-FEATURES model, *telephone*N is replaced by the semantic features associated with *telephone*2 (three SFS) and in the RELATED-WORDS model the lemma is replaced by *telephone*2 and all word senses that are related to it following the definition of relatedness given in Chapter (6.3.4), viz. *call*1, *conversation*2 and *phone*1).

In other words, this one context word yields:

- one context feature in the W model;
- three individual context features in the SF model;

- four individual context features in the REL-W model.

Since the present experiments are based on a set of words where all words are sense-tagged, the counts for each context feature in the two Mirrors-derived models will always be *at least the count that is also found in the corresponding WORD model*, because:

- The W context feature only reflects how many times this specific context feature occurred;
- The SF context features specifies how many words in the training material shared each individual context feature (for instance a very general SF will have a very high count);
- The count of a REL-W context feature is the summed number of times that any member of a group of related word senses occurred in the context.

The RELATED-WORD count means, specifically, that if the word sense *telephone*2 and the word sense *phone*1 (which are related, according to our definition of RELATED-WORDS) were actually seen four times each in the context, each of the two context features will get a frequency of eight; the same frequency will be recorded for the two context features *call*1 and *conversation*2, even if neither of these word senses were actually observed in the context of *bill*N.

Based on the theoretical discussion of the knowledge sources in Chapter (6) and on the first set of experiments in Chapter (9), our observations so far indicate that we may expect the SEMANTIC-FEATURES to be too general to be useful, and in this respect the most interesting comparison will probably concern WORDS and RELATED-WORDS.

Concretely, as specified in Chapter (6.4.2), the theoretical 'worst case' should then be that there is no difference between a WORD-based and a RELATED-WORD-based classification (at least when regarding the outcome within the same context window); in the best case there could be a gain because the REL-W-based classifier learns all the actually occurring context words (although seen as a word sense) and in addition it learns all the word senses that are related to the actually occurring context word senses. If the Mirrors-derived semantic information results in a performance loss, however, it will be very interesting to consider more closely those instances that came out less well in the RELATED-WORD-based model. Specifically, if the use of RELATED-WORDS give poorer result than the context lemmas themselves, two explanations are conceivable: either there is a reason to question the plausibility of the Mirrors-derived information, or there is a reason to question the assumption that if one word sense can co-occur with a target word, then so can also its semantic relatives (even when these are intuitively plausible as semantic relatives).

The remainder of this chapter is organised as follows: Section (10.2) presents the results from the model selection experiments, evaluated with five-fold cross validation. Section (10.3) presents the results when applying each of the three knowledge sources in isolation (WORDS, SEMANTIC-FEATURES and RELATED-WORDS) on the same held-out data sets. Based on the model selection in the preceding section, the best model of each target word is used. Section (10.4) presents the results of EXP8, in which the sense distinctions predicted by the Mirrors method is removed for each context word in order to test to what extent useful relations between context words exist not only within the Mirrors sense predicted through automatic sense-tagging. We close with a discussion and a conclusion in Section (10.5).

## 10.2 Model selection

### 10.2.1 Results cross-validation

The model selection phase has been motivated and outlined in some detail in Chapter (9.2); the model selection of the experiments in the current chapters are performed exactly in the same way. The development results in Chapter (9) were commented in some detail in order to also indicate how to read and interpret the tables. Given the background in the previous chapter, the development results of the current chapter will therefore be less 'walked through' in detail.

So as in Chapter (9), the full tables for the model selection experiments of each target word are listed alphabetically in Appendices 6–8. Appendix 6 shows the cross validation results for EXP5 (WORDS based on the $n$ nearest *sense-tagged* open-class words), Appendix 7 for EXP6 (SEMANTIC-FEATURES), and Appendix 8 for EXP7 (RELATED-WORDS). Each appendix has one table per target word, showing the overall accuracy (recall) for each of the 81 classifiers. The baseline is computed as the relative frequency of the *most frequent sense* (MFS) in the total development data set, and is given above each table. The best classifier in each of the nine window size categories is marked in bold.

#### General analysis of the cross validation results

As in Chapter (9), the best classifier for each target word from cross-validation is summed up in three tables below: WORDS in Table (10.1) (p. 239), SEMANTIC-FEATURES in Table (10.2) (p. 239) and RELATED-WORDS in Table (10.3) (p. 240) (the tables are explained on p. 202).

As a brief summary of the WORD model in Table (10.1), the best models are found in the medium and narrow context window categories. Specifically,

Table 10.1: The best classifier from cross-validation: EXP5: WORDS

| TW | MFS(%) | $naivebayes(l, r)$ | category | R (%) |
|---|---|---|---|---|
| *frisk*AJ | 68.3 | (1 . 30) | *NM* | 70.7 |
| *full*AJ | 94.1 | (4,1) | *NN* | 86.1 |
| *fyr*N | 78.8 | (2,4) | *NN* | 84.2 |
| *gal*AJ | 77.6 | (20,20) | *MM* | 79.3 |
| *lag*N | 70.3 | (30,4) | *MN* | 94.6 |
| *liv*N | 98.1 | (4,1) | *NN* | 93.8 |
| *plan*N | 87.2 | (2,10) | *NM* | 90.8 |
| *rot*N | 80.4 | (30,20) | *MM* | 87.5 |
| *slag*N | 55.6 | (4,30) | *NM* | 55.6 |
| *stemme*N | 92.2 | (20,2) | *MN* | 99.4 |
| *tak*N | 47.5 | (2,4) | *NN* | 65.4 |
| *trykke*V | 80.4 | (50,4) | *WN* | 87.0 |
| *utsette*V | 67.5 | (10,10) | *MM* | 76.6 |
| *utvalg*N | 60.9 | (1,75) | *NW* | 97.8 |
| *valg*N | 60.6 | (30,10) | *MM* | 92.3 |

the complete tables for all target words (Appendix 6) indicate that the better-performing models of a target word are generally clustered in and around the medium context windows, whereas the extremes in both the narrow and wide direction fare less well. All the target words reach or beat the baseline at least once, except *full*AJ and *liv*N, as was also seen with the WORD models in EXP1 in Chapter (9).

Table 10.2: The best classifier from cross-validation: EXP6: SEMANTIC-FEATURES

| TW | MFS(%) | $naivebayes(l, r)$ | category | R (%) |
|---|---|---|---|---|
| *frisk*AJ | 68.3 | (2,1) | *NN* | 62.2 |
| *full*AJ | 94.1 | (1,2) | *NN* | 73.2 |
| *fyr*N | 78.8 | (20,2) | *MN* | 82.5 |
| *gal*AJ | 77.6 | (4,2) | *NN* | 74.1 |
| *lag*N | 70.3 | (10,1) | *MN* | 86.5 |
| *liv*N | 98.1 | (1,2) | *NN* | 92.7 |
| *plan*N | 87.2 | (2,2) | *NN* | 80.7 |
| *rot*N | 80.4 | (2,20) | *NM* | 82.1 |
| *slag*N | 55.6 | (2,20) | *NM* | 45.1 |
| *stemme*N | 92.2 | (4,2) | *NN* | 97.6 |
| *tak*N | 47.5 | (4,1) | *NN* | 59.5 |
| *trykke*V | 80.4 | (20,4) | *MN* | 80.4 |
| *utsette*V | 67.5 | (4,1) | *NN* | 71.4 |
| *utvalg*N | 60.9 | (10,10) | *MM* | 89.1 |
| *valg*N | 60.6 | (10,2) | *MN* | 81.7 |

As for the SEMANTIC-FEATURES in EXP6, Table (10.2) shows that the narrow and medium window size categories once again dominate. A comparison against the baseline indicates that the SEMANTIC-FEATURE classifiers generally have a lower performance than the WORD classifiers in EXP5, as only eight of the fifteen target words have at least one development model that reaches or beats the

baseline.

Table 10.3: The best classifier from cross-validation: EXP7: RELATED-WORDS

| TW | MFS(%) | $naivebayes(l,r)$ | R (%) |
|---|---|---|---|
| *frisk*AJ | 68.3 | ((2,30) , 10) | 70.7 |
| *full*AJ | 94.1 | ((4,1) , 20) | 86.8 |
| *fyr*N | 78.8 | ((20,10) , 10) | 86.0 |
| *gal*AJ | 77.6 | ((10,1) , 40) | 75.9 |
| *lag*N | 70.3 | ((30,10) , 10) | 94.6 |
| *liv*N | 98.1 | ((1,2) , 40) | 95.2 |
| *plan*N | 87.2 | ((2,10) , 4) | 89.9 |
| *rot*N | 80.4 | ((1,30) , 30) | 92.9 |
| *slag*N | 55.6 | ((2,30) , 40) | 51.1 |
| *stemme*N | 92.2 | ((10,10) , 15) | 98.2 |
| *tak*N | 47.5 | ((2,2) , 4) | 65.0 |
| *trykke*V | 80.4 | ((50,20) , 40) | 84.8 |
| *utsette*V | 67.5 | ((4,2) , 10) | 80.5 |
| *utvalg*N | 60.9 | ((20,10) , 4) | 95.7 |
| *valg*N | 60.6 | ((10,10) , 50) | 92.3 |

Considering, finally, the RELATED-WORDS[1], the tables essentially show the same pattern as was also seen in the development experiments of the previous chapter. The classifiers based on RELATED-WORDS generally outperform the SEMANTIC-FEATURE-based classifiers, with eleven of the fifteen target words having classifiers equal to or exceeding baseline. Two of these, *full*AJ and *liv*V, did not reach the baseline with WORDS or SEMANTIC-FEATURES, either. The last two target words that never reached the baseline with the RELATED-WORD classifier was *gal*AJ (which was too difficult in the development experiments of the previous chapter, too) and *slag*N. As regards the SynsetLimit, once again there seem to be no clear tendencies as to the best setting for the SynsetLimit value (cf. also Chapter (9)).

## 10.3 Evaluating on held-out data sets

In this section we present the result of applying each of the three knowledge sources in isolation (WORDS, SEMANTIC-FEATURES and RELATED-WORDS) on the same held-out data sets. Based on the model selection in the preceding section, the best model of each target word is used. This is the same data set as used in Chapter (9), so the sum of test instances across all target words is 1219; each target

[1]Recall from Chapter (9) that the window category column is removed in the RELATED-WORDS table, since the cross-validation experiments combine the nine best models from the WORD experiments and nine values of the SynsetLimit, so only one of the dimensions pertain to window categories.

word has on average 80 test instances and the minimum and maximum numbers of test instances are 16 and 397, respectively.

The classification outcomes will be analysed from three points of view:

- Comparing each knowledge source against the baseline (the most frequent sense, MFS).
- Comparing the two Mirrors-derived knowledge sources against the 'traditional' WORDS classifier, the latter representing the 'best-known' approach.
- Analysing whether the classifiers make complementary errors, in particular studying those instances that were wrongly classified in the RELATED-WORDS model and correctly classified with the counterpart WORD model (the concept of 'WORD counterparts' are explained in Chapter (9.3.1)).

### 10.3.1 Evaluating the knowledge sources individually

**Overview**

The results of testing each classifier on the same data sets are summed up in Table (10.4) (p. 242). The target words are listed in the first column and the second column lists the baselines per target word (the number of test instances that would be correctly classified simply by choosing the most frequent sense—the MFS—from training). The last three columns show the overall accuracy (recall) when applying each knowledge source individually. The overall accuracy is measured in absolute numbers (in the column marked with a #) and with the corresponding ratio (in the column marked with %). Below the fifteen target word results in the table is the sum of correct classifications totally per knowledge source across all target words.

Based on the total counts at the bottom of the table, the overall best knowledge source for classification is RELATED-WORDS, which has *one* more correct classification than WORDS (1050 correct against 1049 correct classifications, respectively). By contrast, the SEMANTIC-FEATURES model performs markedly worse with a performance decrease from .861 (WORDS and RELATED-WORDS) to .808. A pairwise comparison of the classification outcomes across all target words reveal that the difference between the WORD-based classifier and the RELATED-WORDS-based classifier is larger than suggested by the overall accuracy—we will return to this in more detail in Section (10.3.4)–but nonetheless the difference is not judged to be statistically significant. The overall pairwise difference between WORDS and SEMANTIC-FEATURES, on the other hand, is statistically very significant (to be discussed more in detail in Section (10.3.3)).

Table 10.4: Overall Accuracy for the individual knowledge sources (#=absolute counts, %=relative proportion

| TW | MFS | WORDS | | SEMANTIC-FEATURES | | RELATED-WORDS | |
|---|---|---|---|---|---|---|---|
| | | # | % | # | % | # | % |
| *frisk*AJ | .500 | 22/36 | .611 | 21/36 | .583 | 27/36 | .750 |
| *full*AJ | .947 | 173/189 | .915 | 142/189 | .751 | 169/189 | .894 |
| *fyr*N | .680 | 20/25 | .800 | 17/25 | .680 | 19/25 | .760 |
| *gal*AJ | .706 | 34/51 | .667 | 39/51 | .765 | 40/51 | .784 |
| *lag*N | .647 | 13/17 | .765 | 14/17 | .824 | 12/17 | .706 |
| *liv*N | .982 | 377/398 | .947 | 371/398 | .932 | 383/398 | .962 |
| *plan*N | .745 | 34/47 | .723 | 33/47 | .702 | 33/47 | .702 |
| *rot*N | .800 | 20/25 | .800 | 16/25 | .640 | 20/25 | .800 |
| *slag*N | .544 | 26/57 | .456 | 23/57 | .404 | 30/57 | .526 |
| *stemme*N | .944 | 143/144 | .993 | 138/144 | .958 | 144/144 | 1.000 |
| *tak*N | .505 | 84/111 | .757 | 72/111 | .649 | 73/111 | .658 |
| *trykke*V | .714 | 17/21 | .810 | 14/21 | .667 | 14/21 | .667 |
| *utsette*V | .697 | 27/33 | .818 | 27/33 | .818 | 28/33 | .848 |
| *utvalg*N | .550 | 20/20 | 1.000 | 19/20 | .950 | 19/20 | .950 |
| *valg*N | .578 | 39/45 | .867 | 39/45 | .867 | 39/45 | .867 |
| Total | | 1049/1219 | .861 | 985/1219 | .808 | 1050/1219 | .861 |

### The knowledge sources vs. the baseline

When comparing each knowledge source against the baseline for each target word, there seems to be no substantial difference between the three knowledge sources. The RELATED-WORDS classifier is marginally better than the other two knowledge sources, reaching or beating the baseline with exactly two thirds of the target words (ten of fifteen target words) whereas the two other knowledge reach or beat the baseline with one target word less (nine of the target words).

### The two Mirrors-derived knowledge sources vs. the WORD model

As for the effect of applying each of the two Mirrors-derived knowledge sources compared to the 'traditional' WORD classifier, Table (10.4) indicates a performance decline from the best WORDS model to the best SEMANTIC-FEATURES model with nine of the target words (almost two thirds). With three of the words the use of SEMANTIC-FEATURES yield a slight improvement and with two target words the overall accuracy is unchanged. Of these, only two observations were judged to be statistically significant ($\alpha = 0.05$), viz. the observed decrease with *full*AJ and *tak*N.

From the best WORDS model to the best RELATED-WORDS model there was a decline with seven target words, an improvement with six of the words and an unchanged accuracy with two target words, and none of these differences were judged to be statistically significant.

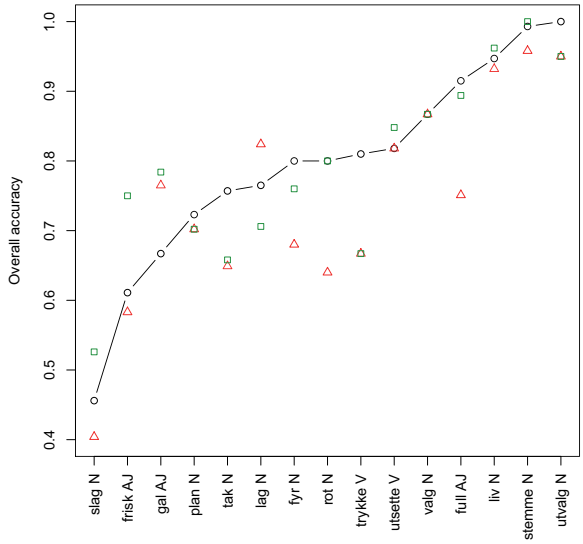



Figure 10.1: Overall Accuracy from Table (10.4) sorted by W. Legend: black=W, red=SF, green=REL-W.

A plot of the overall accuracy (OA) from Table (10.4) (p. 242) sorted by increasing OA for WORDS is depicted in Figure (10.1). The figure shows that for each word, there are no extreme OA differences between the models although the SF (red circles) accuracies are generally slightly lower than with WORDS (the black line) and the RELATED-WORDS (green circles).

### Summing up

Summing up on the basis of the overall accuracies (OAs), the experiments of the present chapter seem to indicate the same pattern that was seen in Chapter (9): the knowledge source SEMANTIC-FEATURES is 'the odd one out' whereas the difference between using semantically unanalysed WORDS and using the RELATED-WORDS is only marginal.

In order to show the span between accuracies in each model, Table (10.2) plots the overall accuracies from Table (10.4) (p. 242) sorted independently for each model. The figure indicates that the W and the REL-W are fairly similar whereas the SF performs slightly worse in terms of OA.

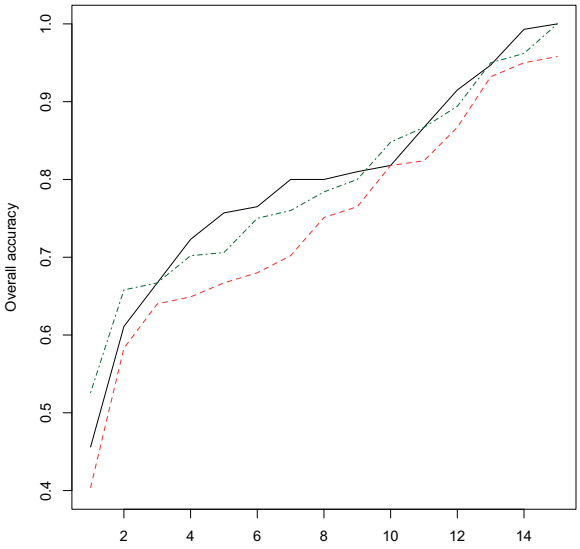



Figure 10.2: Overall Accuracy from Table (10.4) sorted by each model independently. Legend: black=W, red=SF, green=REL-W.

It may be remarked that the statistical analysis in relation to SEMANTIC-FEATURES is contradictory in that differences of classification outcomes, when measured across all target words taken together (i.e. with a stronger statistical basis), are found to be statistically significant whereas the statistical significance findings diminish when judging the classification outcomes of each target word individually. Statistically speaking, the probability of obtaining values indicative of a large (significant) difference by chance increases with the decrease in the sample size, which could be taken to motivate that we pay greater heed to the findings based on the largest sample size (when viewing all classifications across the target words). It seems, at any rate, clear that the difference between SEMANTIC-FEATURES and WORDS is greater than the difference between RELATED-WORDS and WORDS. In the following sections each of the knowledge sources will be given a formal analysis in terms of the training models and classification outcomes (Sections 10.3.2-10.3.4).

### 10.3.2 EXP5: The WORD classifier

Table 10.5: Context features during training and testing: EXP5 (WORDS)

| | | **Training context features** | | | **Test lemmas** | | |
|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Known | Total | Ratio |
| *frisk*AJ | (1,30) | 1135 | 2.231 | 1.000 | 249 | 687 | .362 |
| *full*AJ | (4,1) | 1074 | 2.044 | 1.000 | 197 | 642 | .307 |
| *fyr*N | (2,4) | 227 | 1.507 | 1.000 | 26 | 118 | .220 |
| *gal*AJ | (20,20) | 1526 | 3.041 | 1.000 | 406 | 1012 | .401 |
| *lag*N | (30,4) | 594 | 2.118 | 1.000 | 112 | 394 | .284 |
| *liv*N | (4,1) | 1699 | 2.723 | 1.000 | 452 | 1048 | .431 |
| *plan*N | (2,10) | 761 | 1.719 | 1.000 | 131 | 411 | .319 |
| *rot*N | (30,20) | 1286 | 2.166 | 1.000 | 276 | 730 | .378 |
| *slag*N | (4,30) | 1800 | 2.509 | 1.000 | 429 | 1147 | .374 |
| *stemme*N | (20,2) | 2065 | 3.547 | 1.000 | 701 | 1477 | .475 |
| *tak*N | (2,4) | 750 | 2.048 | 1.000 | 167 | 452 | .369 |
| *trykke*V | (50,4) | 1081 | 2.298 | 1.000 | 270 | 715 | .378 |
| *utsette*V | (10,10) | 779 | 1.977 | 1.000 | 138 | 447 | .309 |
| *utvalg*N | (1,75) | 986 | 3.452 | 1.000 | 321 | 894 | .359 |
| *valg*N | (30,10) | 1337 | 3.111 | 1.000 | 418 | 923 | .453 |
| Averaged ratio of known test context features across target words: | | | | | | | .361 |
| (standard deviation: | | | | | | | $\pm$.067) |

Table (10.5) shows the relation between context features during training and during testing with the best WORDS model. Columns three, four and five show that the result from training can be compared with the outcome in EXP1 in Chapter (9): the average frequency counts per context feature (per lemma registered in the training model) are around 2 (the maximum average value is $3.547$ and the minimum average value is $1.507$) and the median for each target word is at hapax

level. Considering the ratio of known test lemmas (the three last columns), the average ratio was roughly a third (the averaged ratio of known context features across target words was .361 $\pm$.067.

### 10.3.3 EXP6: The SEMANTIC-FEATURE classifier

Considering SEMANTIC-FEATURES, each context feature now represents a semantic feature seen in the context. Our starting point from the overview is that there is a decrease in performance when comparing the best WORD model against the best SEMANTIC-FEATURES model (Section (9.3.2)). This section will look into the reasons for this in some detail.

We will begin with the training model. If the SEMANTIC-FEATURES did have a positive effect compared to the use of unanalysed WORDS, this ought to be reflected in higher frequency counts (because unanalysed words are now counted together). However, Table (10.6) indicates that the average frequency across all context features of a target word actually remains quite low; the same is indicated by the median for each target word, which consistently equals 1 for the context feature counts with all target words. This strongly suggests that the correspondingly low frequency counts for SFs in EXP2 (Chapter (9)) did *not* hinge on a trade-off between adding Mirrors information and losing context information.

Table 10.6: Context features during training and testing: EXP6 (SEMANTIC-FEATURES)

| | | **Training context features** | | | **Test lemmas** | | |
|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Known | Total | Ratio |
| *frisk*AJ | (2,1) | 306 | 2.314 | 1.000 | 50 | 80 | .625 |
| *full*AJ | (1,2) | 1000 | 3.653 | 1.000 | 204 | 410 | .498 |
| *fyr*N | (20,2) | 917 | 4.726 | 1.000 | 213 | 376 | .566 |
| *gal*AJ | (4,2) | 579 | 4.655 | 1.000 | 113 | 190 | .595 |
| *lag*N | (10,1) | 432 | 2.792 | 1.000 | 73 | 151 | .483 |
| *liv*N | (1,2) | 1556 | 5.983 | 1.000 | 417 | 670 | .622 |
| *plan*N | (2,2) | 530 | 2.645 | 1.000 | 86 | 147 | .585 |
| *rot*N | (2,20) | 947 | 4.147 | 1.000 | 198 | 369 | .537 |
| *slag*N | (2,20) | 1716 | 5.637 | 1.000 | 430 | 825 | .521 |
| *stemme*N | (4,2) | 1153 | 6.677 | 1.000 | 300 | 507 | .592 |
| *tak*N | (4,1) | 906 | 5.039 | 1.000 | 198 | 368 | .538 |
| *trykke*V | (20,4) | 887 | 4.081 | 1.000 | 188 | 358 | .525 |
| *utsette*V | (4,1) | 424 | 2.545 | 1.000 | 66 | 114 | .579 |
| *utvalg*N | (10,10) | 654 | 4.231 | 1.000 | 166 | 305 | .544 |
| *valg*N | (10,2) | 896 | 4.051 | 1.000 | 215 | 377 | .570 |
| Averaged ratio of known test context features across target words: | | | | | | | .559 |
| (standard deviation: | | | | | | | $\pm$.042) |

In order to analyse this seeming lack of generalisation more closely, we will remove the dependence of the best W model and the best SF model on different context window sizes, and we will consider the best SEMANTIC-FEATURES classifiers for each target word and its counterpart WORD model. The overall counts

of similar and changed classifications across all target words are given in the contingency table in Table (10.7). The truth values denote correct (true) and wrong (false) classifications. Viewing all classifications across all fifteen target words, the SF model has an OA of .808, and is thus outperformed by its counterpart W which has an OA of .824 (the former is found in Table (10.4) (p. 242) and the latter may be derived from the contingency table in Table (10.7) by summing the 'True' counts for W, i.e. $906 + 98$, divided by the total, 1219). Whereas the difference between the best W model and the best SF model (cf. Table (10.4)) was found to be statistically significant, the best SFs does not show a statistically significant decrease in performance compared to its counterpart W, based on McNemar's test.

| | | **SF** | |
|---|---|---|---|
| | | True | False |
| **W** | True | 906 | 98 |
| | False | 79 | 136 |

Table 10.7: A pairwise comparison between the SF model (EXP5) and its WORD counterparts for all target words).

So why is there such a small difference between the best SF model and its counterpart W? An analysis of the learning models in the two models–the best SF model and its counterpart W model–showed that as one would expect, the frequency distribution in a WORD-based model resembles the so-called Zipfian distribution (cf. Chapter (3.5.1)): a few context features are very frequent and many context features are extremely low-frequent. This tendency is naturally stronger in the case of target words with small context windows and few examples, but the tendency appears to be consistent across target words, including those with higher frequencies. It would take this chapter a bit too far to illustrate with detailed curves for every target word; instead the general point is illustrated by way of *one* target word example (the target word *frisk*AJ): the curves in Figure (10.3) (p. 248) show the frequency curves of the context features observed in the W model (to the left) which is the counterpart W to the best SF-based model (to the right). Ideally one would hope that by abstracting to SEMANTIC-FEATURES, one might alleviate this extreme drop from high- to low-frequent context features because singular words may be grouped together in virtue of sharing SFs. Considering the model generalisations from the counterpart W to the best SF, however, the following two points seem to be a general tendency:

- a few SEMANTIC-FEATURES grouped many context word senses (the frequencies, seen on the vertical axis in the figures, are higher in the SF models than in the counterpart W models, not only in the *frisk*AJ example) whereas many SFs did not at all generalise across context words (i.e. in most cases

only *one* context word gave rise to the registration of a given SF in the trained model), and

- the resulting context features in the SF models have a similar Zipf-like distribution to that seen with a W-based model.

This probably relates to the inescapable data sparseness problem. It was known beforehand that many semantic fields are quite small in the Mirrors, but since the Mirrors is an *abstraction* from the ENPC, it was not known beforehand how the frequencies would come out when mapping the abstract Mirrors information back to the corpus (through automated sense-tagging). Sparse data means, in the case of the Mirrors method, that one must expect a good many lexical gaps in the corpus data given as input to the Mirrors method. Lexical gaps, in turn, may cause intuitively related words to be placed in separate semantic fields because no direct or indirect translational overlap was found in the corpus. For this reason it tends to be the case that the Mirrors method finds relatively few related words and hence, many SFs are only supported in the context by the word that gave rise to it. Unfortunately, this must be attributed to the corpus size and not to properties of the Mirrors as such, which means that it becomes difficult to make claims about the representativity of the presented observations.

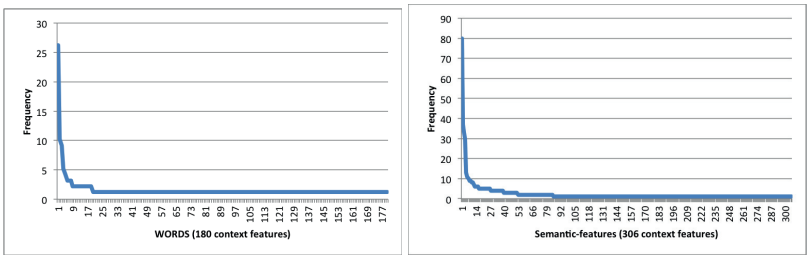



Figure 10.3: A Zipf-like frequency distribution of the context features in the best W-based (left-hand curve) and the best SF model (right-hand curve) for the target word *frisk*AJ (EXP5, EXP6).

So the frequency distribution, as they come out with the given data material, suggests that the abstraction from Ws to SFs does not lead to a more compact model as one would perhaps hope. Moreover, the fact that the classification accuracy decreases with the use of SEMANTIC-FEATURES–although the direct comparison between the best SFs and their W counterparts do not yield a statistically *significant* decrease–indicate that the generalisations are not unequivocally positive for WSD. This motivates a brief study of some of the generalisations found by the Mirrors method in the contexts of the target words.

The interesting property of the trained models, in this respect, is that they visualise, as it were, the relationship between (i) words found to be related according to the *Mirrors method* and (ii) words that in fact occur in similar contexts according to the corpus (although the small data samples available in the current thesis prohibit strong claims about the generality of conclusions). Recall the 'learnability' criterion in Chapter (5.3) (p. 97). A basic assumption of our experiments is that we expect that the usage of a given word hinges on discoverable regularities. If the translation-based (Mirrors-derived) senses and relations are retrieved in corpus data, then we can point to a correlation between the translational sense criterion of the Mirrors and the context-based sense criterion. The trained models are thus interesting since a SF, by itself, has a fixed set of word senses in its denotation, whereas the trained models will visualise subsets of word senses where the translation-based criterion (the Mirrors method) and the context-base sense criterion overlap.

It was not easy to find a straightforward way to identify whether an awkward link between words in the Mirrors pertains to word alignment errors or to the assumptions of the Mirrors method. No attempt was therefore made to do a systematic evaluation manually across all target word models. A simple, manual study of the trained SF models suggests the following trends:

First of all, the most general SFs found in a model tend to contain the least plausible groupings. Considering for instance the SF model for *frisk*AJ, the most frequently seen SEMANTIC-FEATURE during training was $[have1|vaere1]$, which had a frequency of $80$ and which united $38$ unique context word senses during training, comprising a variety of context lemmas spanning from *være*V to *våkne*V ('wake up'), *brenne*V ('burn') and *starte*V ('start').

As regards the smaller groups of generalisations found in the target word contexts, these are more often plausible. To illustrate the kinds of connections that are found, some examples are given in Tables 10.8–10.10 (the English translations are given by the author). The listed examples were all taken from the best SF model for *slag*N.

When studying the full material in the Mirrors word bases it tends to be the case that noun senses are found in very small and relatively 'flat' semantic fields whereas adjectives and verbs tend to have more depth in their semantic hierarchies. Considering the generalisations found in the training material with SFs, many good connections between nouns were found, although they were usually restricted to two or three nouns (cf. Table (10.8)). For instance, the first row in Table (10.8) indicates that whereas the three lemmas in the right-hand column would be counted individually in a W model, their presence is recorded and counted jointly in the SF model because they share a semantic feature. With adjectives (cf. Table (10.9)), the present analysis has not been performed sufficiently in detail to make strong claims, but this is the word class where good examples of plaus-

Table 10.8: Ten SFs s grouping context nouns during training.

| | |
|---|---|
| $[door2\|doer1]$ | *dør* 'door', *døråpning* 'door opening', *kirkedør* 'church door'. |
| $[bedroom1\|rom1]$ | *rom* 'room', *soverom* 'sleeping room'. |
| $[spray1\|sprut1-*]$ | *spray* 'spray', *bedøvelsesspray* 'anaesthetic spray'. |
| $[talk1\|grad1]$ | *samtale* 'conversation', *prat* 'chat'. |
| $[while1\|time1]$ | *stund* 'while', *time* 'hour'. |
| $[occasion1\|situasjon1]$ | *begivenhet* 'event', *anledning* 'occasion', *Kongen* 'king'. |
| $[worker2\|arbeidstaker1]$ | *arbeider* 'worker', *arbeidstaker* 'worker'. |
| $[war1\|strid1]$ | *strid* 'combat', *krig* 'war'. |
| $[knowledge2\|tank1]$ | *erfaring* 'experience', *kunnskap* 'knowledge', *opplysning* 'information'. |
| $[characteristic2\|kjennetegn1]$ | *egenskap* 'property', *kvalitet* 'quality'. |

Table 10.9: Ten SFs s grouping context adjectives during training.

| | |
|---|---|
| $[peculiar1\|merkelig1]$ | *merkverdig* 'odd', *merkelig* 'odd'. |
| $[lovely1\|vakker1]$ | *fantastisk* 'fantastic', *vakker* 'beautiful', *koselig* 'nice', *flott* 'great'. |
| $[small1\|spinkel1]$ | *lav* 'low', *beskjeden* 'modest', *personlig* 'personal', *lite* 'little'. |
| $[wide1\|bred1]$ | *vid* 'wide', *bred* 'broad'. |
| $[late1\|gammel2]$ | *ofte* 'often', *lenge* 'long', *sist* 'last', *sen* 'late', *gammel* 'old'. |
| $[nice1\|god1]$ | *vakker* 'beautiful', *koselig* 'cosy', *pen* 'pretty', *fin* 'nice', *god* 'good', *ny* 'new', *mye* 'much'. |
| $[exceptional1\|fabelaktig1]$ | *eventyrlig* 'fabulous', *enestående* 'outstanding'. |
| $[hard1\|fast1]$ | *tung* 'heavy', *høy* 'tall', *kort* 'short'. |
| $[tight1\|tett1]$ | *riktig* 'right/correct', *dårlig* 'bad', *lite* 'little'. |
| $[able1\|mye1]$ | *full* 'full', *lenge* 'long', *mye* 'much', *lett* 'little'. |

ibly connected context words were hardest to find. This is interesting since the typical semantic vagueness of many adjectives usually means that much translational overlap is found, yielding often many as well as plausible relations between adjective word senses. In the context of WSD, however, it may be that the same properties lead to too general information. As for verbs, this was perhaps the word class where it was easiest to find nice examples of classes of semantically related words (cf. Table (10.10)).

Regarding, finally, known and unknown lemmas as classification time, Figure (10.4) uses a 100% stacked bar chart to show the distribution of known and unknown test lemmas in the best SF models and their counterpart Ws. As in the

Table 10.10: Ten SFs grouping context verbs during training.

| | |
|---|---|
| $[choose1\|oenske1]$ | *foretrekke* 'prefer', *ønske* 'wish' (but also *spille* 'play'). |
| $[maintain1\|hevde1]$ | *påstå* 'contend', *hevde* 'claim'. |
| $[direct1\|feste1]$ | *lede* 'lead', *kontrollere* 'control', *styre* 'govern,steer'. |
| $[gain1\|oppnaa]$ | *vinne* 'win', *oppnå* 'gain'. |
| $[notice1\|merke1]$ | *fornemme* 'sense', *oppdage* 'discover', *merke* 'sense'. |
| $[occur1\|hende1]$ | *hende* 'happen', *forekomme* 'occur', *skje* 'happen'. |
| $[argue1\|forklare1]$ | *diskutere* 'discuss', *forklare* 'explain', *hevde* 'claim'. |
| $[struggle1\|proeve1]$ | *forsøke* 'try', *prøve* 'try', *kjempe* 'fight'. |
| $[sense1\|skjoenne1]$ | *fornemme* 'semse', *føle* 'feel', *skjønne* 'understand'. |
| $[occur1\|hende1]$ | *hende* 'happen', *forekomme* 'occur', *skje* 'happen'. |

corresponding figures in Chapter (9), each bar represents one target word. As can be seen, the two knowledge sources appear to be more similar than different; most test lemmas are known in both models ($TT$) or unknown in both models ($FF$). Since only sense-tagged context words were considered, everything that is known in the W model is also known in the Mirrors-derived model and therefore the $TF$ counts equal zero. Approximately a third of the test lemmas are only known to the SF model, although this did not lead to any improvement in accuracy.

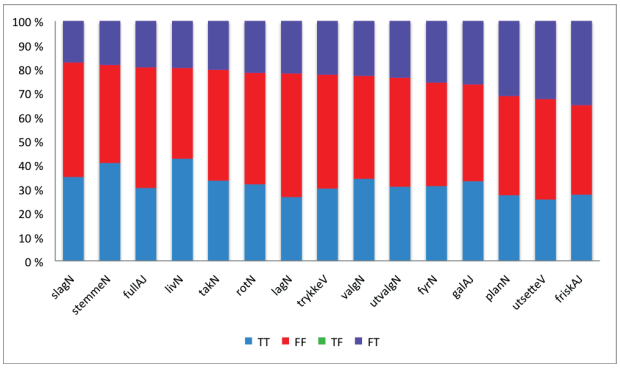



Figure 10.4: The distribution of known and unknown test lemmas in the best SF model per target word vs. the counterpart W.

Summing up on SEMANTIC-FEATURES, the results suggest that irrespective of whether the SF-based generalisations are plausible or not, they seem to have the same unwanted properties that we also find in a traditional W model: some SFs are very frequent (and the most frequent SFs then have a markedly higher frequency than the most frequent context features in a W model) whereas *most* SFs are actually extremely low-frequent. Comparing the classification outputs, the W and SF models largely classify on the basis of the same set of known context lemmas.

### 10.3.4 EXP7: The RELATED-WORDS classifier

Our starting point from the overview in Section (10.3.1) is that there is no statistically significant difference between WORDS-based classification and classification based on REL-WS when using the best model from model selection with both knowledge sources. The same finding is made when comparing the best RELATED-WORDS model against its counterpart W, i.e. when all factors are equal except the knowledge source. This latter comparison is summed up in Table (10.11), which sums up the counts of a pairwise comparison per test instance across all target words between the two models. Of the total data set, 54 instances were correctly

classified by the REL-W and wrongly classified with the W model; the other way around the W was correct and the REL-W was wrong 42 times (so in other words, the REL-W was slightly better than the W classifier with a precision of .861 against .852, although not significantly better).

| | | **REL-W** | |
|---|---|---|---|
| | | True | False |
| **W** | True | 996 | 42 |
| | False | 54 | 127 |

Table 10.11: A pairwise comparison between the SF model (EXP7) and its WORD counterpart for each target word.

It is interesting to note that the best WORD model and the best RELATED-WORDS model are not found to have statistically significant differences in spite of the fact that the best WORDS models build on roughly half as many context features during learning as the best RELATED-WORDS models. This can be seen from Table (10.12) (p. 253), which shows some counts from the training model and the classification analysis for each target word. Comparing the numbers of context features in column three of this table with the corresponding columns in Table (10.5) (p. 245), the figures show that the best REL-W model consistently learns roughly twice as many context features as the best W model. Considering the ratio of known lemmas at classification time (the last three columns of the table), the ratio of known lemmas is slightly higher in the REL-W model compared to the best W model in Table (10.5): the average ratio of known lemmas is .498 in in the former model and .361 in the latter.

Their differences at classification time do not however appear marked at all. This is conveniently visualised in a 100% stacked bar chart showing the distribution of known and unknown test lemmas in the best RELATED-WORDS model and its counterpart W (Figure (10.5)). Given precisely the same context lemmas available, the figure reveals that in the presented experiments between 80% and 90% the context lemmas at classification time are either known (labelled $TT$ in the figure) or unknown (labelled $FF$) in both models. The last 10–20% are only known in the REL-W model (the $FT$ lemmas). In other words, the lack of a difference between the models in terms of OA seems to find its primary explanation in the fact that the Mirrors RELATED-WORDS did not seem to add very often extra knowledge *that turned out to be useful at classification time*.

A scrutiny of the instances where the REL-W classifier and the W classifier make diverging classification choices does not indicate that there are very general circumstances that can account for the cases where they make different classification choices. For instance there is no indication that the examples of success

Table 10.12: Context features during training and testing: EXP7 (RELATED-WORDS)

| | | **Training context features** | | | **Test lemmas** | | |
|---|---|---|---|---|---|---|---|
| TW | model | context features | average-N | median | Known | Total | Ratio |
| *frisk*AJ | ((2-30) , 10) | 3027 | 3.404 | 1.000 | 346 | 700 | .494 |
| *full*AJ | ((4-1) , 20) | 2759 | 2.993 | 1.000 | 294 | 642 | .458 |
| *fyr*N | ((20-10) , 10) | 2206 | 3.077 | 1.000 | 238 | 489 | .487 |
| *gal*AJ | ((10-1) , 40) | 1774 | 2.917 | 1.000 | 167 | 349 | .479 |
| *lag*N | ((30-10) , 10) | 1965 | 2.838 | 2.000 | 203 | 455 | .446 |
| *liv*N | ((1-2) , 40) | 2785 | 4.040 | 2.000 | 383 | 670 | .572 |
| *plan*N | ((2-10) , 4) | 2129 | 2.365 | 1.000 | 199 | 411 | .484 |
| *rot*N | ((1-30) , 30) | 2448 | 2.811 | 1.000 | 237 | 491 | .483 |
| *slag*N | ((2-30) , 40) | 3945 | 4.364 | 2.000 | 543 | 1090 | .498 |
| *stemme*N | ((10-10) , 15) | 4083 | 6.742 | 2.000 | 788 | 1343 | .587 |
| *tak*N | ((2-2) , 4) | 1669 | 2.458 | 1.000 | 148 | 313 | .473 |
| *trykke*V | ((50-20) , 40) | 3092 | 4.067 | 2.000 | 450 | 877 | .513 |
| *utsette*V | ((4-2) , 10) | 1095 | 1.699 | 1.000 | 62 | 138 | .449 |
| *utvalg*N | ((20-10) , 4) | 1710 | 3.311 | 2.000 | 206 | 428 | .481 |
| *valg*N | ((10-10) , 50) | 2329 | 3.350 | 2.000 | 314 | 550 | .571 |
| | | | | | Averaged ratio of known test context features across target words: | | ..498 |
| | | | | | (standard deviation: | | ±.044) |

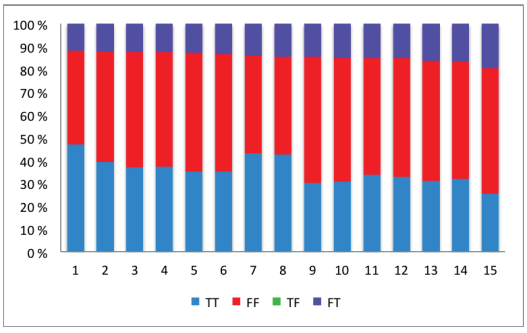



Figure 10.5: The distribution of known and unknown test lemmas in the best REL-W model per target word vs. the counterpart W.

or failure are clearly connected to certain word classes, nor that it can usually be traced back to errors in the Mirrors method information (or in word alignment errors). It rather appears to be the case that sometimes the REL-W model happens to know significant context words which made the classifications go in the right direction; sometimes one does find traces of noise in the Mirrors-derived context information and sometimes the frequencies are simply very different in two models, even for the context information that is known in both models. It is in general extremely difficult to generalise about the reasons why a classification went wrong (or well) because the limited data sets do not enable us to ascertain to what extent the classification outcomes are accidental.

Four examples may illustrate the span in reasons why the classifiers sometimes diverge.

**example1: *rot*N**

An instance of the target word *rot*N was not classified correctly by the W model whereas the REL-W model did choose the right sense. The instance should be classified as representing the 'root' meaning (vs. its 'mess' meaning). The REL-W model succeeded because it knew decisive words such as *terreng* 'terrain', which did not in itself occur during training but which was recorded because *jord* 'earth' was observed with the 'root' meaning several times during training.

**example2: *slag*N**

An instance of the 'sort, type' sense of the target word *slag*N (*slag*1) was correctly recognised by the REL-W in a test instance where the W failed. The target word occurred in the following context:

*..av noe slag..* '..of any sort..'.

The most contributing lemma in the REL-W model was the verb *blande* 'mix', which was unknown in the W model. This lemma was attested seven times in the REL-W model, and only in the context of *slag*1 (as in *blande ting av ulike* ***slag*** 'mixing things of different **sorts**'), but not because this verb itself actually appeared during training. This verb sense became registered during learning because it was found to in a class of RELATED-WORDS with the verbs *bygge* 'build', *studere* 'study' and *regne* 'estimate'. Although one may argue that all of them are more or less connected to construction work, the relation between them is quite loose (to the extent that one cannot even know for sure if there are word alignment errors among them). So whereas it was very suitable that *blande* 'mix' should contribute to the correct classification of this instance, it is not linguistically obvious that a sense of this verb should have been grouped as a RELATED-WORD to the other verbs listed.

**example3: *tak*N**

The target word *tak*N illustrates that sometimes it is the sense divisions of the target word that pose problems in that the difference between senses may be quite small. One instance of *tak*N, where the correct sense is *tak*7 ('roof') was correctly classified by the W model and erroneously classified by the REL-W model.

The local context is as follows: *Inne var det jordgulv, og taket besto av halm eller bølgeblikk.* 'inside there were mud floors and thatched or tin roofs'. The W model chose the correct sense based on the lemma *være* 'be' whereas the REL-W model chose the sense denoting the inside 'ceiling', *tak*4. This choice was made based on three known lemmas; the two that were not known from the actually occurring words during training was *jordgulv* 'mud floor' (which was recorded during training as a relative of *jordgulv* 'mud floor' and *bestå* 'consist'. Since the 'mud floor' concept was associated with *indoor* descriptions during training, the REL-Wchose the 'ceiling' and not the 'roof' sense.

**example4: *liv*N**

The last example will illustrate the problematic aspect that there are errors in the Mirrors material which leads to unexpected generalisations or conclusions, although there is no basis of assessing the extent to which this is a *typical* error. The 'waist' sense of *liv*N, *liv*8, occurred in one of its instances in the context: *(..) la armen om livet (..)* '(..) threw his arm around my waist (..)'. In the W model, the co-occurrence *arm* was quite indicative of the correct sense, *liv*8, in terms of its ratio, occurring twice with *liv*8 and never with *liv*1. In the REL-W model, on the other hand, the frequencies changed radically: The word sense *arm*1 was registered 27 times with *liv*1 and four times with *liv*8. The reason for this is that *arm*1 has been grouped into a quite large semantic field consisting of several subfields where there are partially sensible relations, but where the total picture is more or less a 'salad bowl'. Eight actually occurring context lemmas gave rise to the frequency of *arm*1, in virtue of having this sense as one of its RELATED-WORDS: *jente* 'girl', *forelder* 'parent', *arm* 'arm', *datter* 'daughter', *hår* 'hair', *mann* 'man' and *i*—-the latter probably resulting from a lemmatisation error (or maybe it represents the letter 'i' as a noun).

Summing up on RELATED-WORDS, the results suggest that RELATED-WORDS produce less classification errors than the more general SFs. But on the other they do not seem to help the classifier to learn *more* context words at classification time than a traditional W model. Comparing the classification outputs, the W and REL-W models largely classify on the basis of the same set of known context lemmas.

## 10.4 Evaluating the quality of the Mirrors sense divisions

This last section of the theoretical experiments aims to evaluate the sense divisions of the Mirrors method (cf. the experimental outline in Chapter (6)).

It has already been asserted that the Mirrors method displays a tendency to generate more sense distinctions than what is intuitively desirable. The motivating question of this section is therefore related to the potential use of the Mirrors method as a lexical resource for WSD: if the Mirrors method does generate too many word senses, does this mean that valuable information is actually lost when confining our attention to 'sense-specific' semantic features or related words for WSD?

Thus, in this final experiment a classifier is trained and tested based on the same material as is used the RELATED-WORDS experiments of EXP7, except that the sense *partitions* of the context lemma are dissolved. Whereas the REL-W model in EXP7 rests on the RELATED-WORDS of the context word sense predicted by the automatic sense-tagger, the UNION model of EXP8 rests on the *union* of RELATED-WORDS associated to each It was chosen to test with the same context window settings as in EXP7; hence one may think of this last classifier as a kind of UNION counterpart to the RELATED-WORDS classifier in EXP7 which includes the RELATED-WORDS of each sense that a context word has, according to the Mirrors method. This means that the classifier only directly comparable to the RELATED-WORDS classifier of EXP7.

Instead of collecting the RELATED-WORDS of that sense predicted by the automatic sense-tagging, the Mirrors-derived RELATED-WORDS information from *all* the possible senses of a context word are considered. If the *union* of semantic information about a context word is more useful than the sense-specific information, this would indicate that too much relevant information is spread across several senses, i.e. that the Mirrors senses are not adequate for WSD purposes.

This experiment could in principle be carried out by considering, not only the *sense-tagged* context words, but any context word that has an entry in the Mirrors word bases; its semantic features in every possible sense could then be retrieved. But it may be interesting to consider this experiment in relation to the corresponding sense-specific information in EXP7; therefore the union of Mirrors-derived information was based on the $n$ nearest words that were sense-tagged (although we then do not confine our attention to the particular sense predicted by the automatic sense-tagger in that particular context).

Table (10.13) shows the overall accuracies per target word. Recall that these results are based on the same training and data sets as used in all other experiments, and it rests on the best model settings for RELATED-WORDS.

Table 10.13: Overall Accuracy when ignoring the Mirrors word sense divisions (#=absolute counts, %=relative proportion

| TW | MFS | UNION RELATED-WORDS | |
|---|---|---|---|
| | | # | % |
| *frisk*AJ | 50.0 | 18/36 | .50 |
| *full*AJ | 94.7 | 110/189 | .582 |
| *fyr*N | 68.0 | 13/25 | .520 |
| *gal*AJ | 70.6 | 25/51 | .490 |
| *lag*N | 64.7 | 7/17 | .412 |
| *liv*N | 98.2 | 311/398 | .781 |
| *plan*N | 74.5 | 25/47 | .532 |
| *rot*N | 80.0 | 8/25 | .320 |
| *slag*N | 54.4 | 5/57 | .088 |
| *stemme*N | 94.4 | 35/144 | .243 |
| *tak*N | 50.5 | 75/111 | .676 |
| *trykke*V | 71.4 | 6/21 | .286 |
| *utsette*V | 69.7 | 28/33 | .848 |
| *utvalg*N | 55.0 | 14/20 | .700 |
| *valg*N | 57.8 | 25/45 | .556 |

Table (10.6) (p. 258) plots the overall accuracies (OAs) of the RELATED-WORDS in the fifth column of Table (10.4) and of the counterpart UNION classifier in Table (10.13), sorted by REL-W. This figure indicates that based on the present material of this thesis, the performance decline is quite striking when attempting to include RELATED-WORDS irrespective of the sense predicted by the automatic sense-tagger.

A brief analysis of the classification output indicates that the difference between the best RELATED-WORDS model and the UNION model is extremely small when considering the distribution of known and unknown lemmas in the two models. Figure (10.7) indicates that just about all lemmas in a test situation are known in both models ($TT$) or unknown in both models ($FF$). Only a marginal amount of test lemmas is known only in the UNION model ($FT$), although, as indicated by the overall results, there is no gain in this abstraction. These findings suggest that even though the Mirrors method may derive several senses for a word, the main information of relatedness must usually be clustered within one of the sense.

It is, however, very difficult to assert whether this reflects any property of the Mirrors method or if it pertains to the small amounts of data, therefore we will not pursue this any further. Suffice it to say that based on the present study, there is no reason to believe that there is an information loss in concentrating on the information predicted by the Mirrors method.

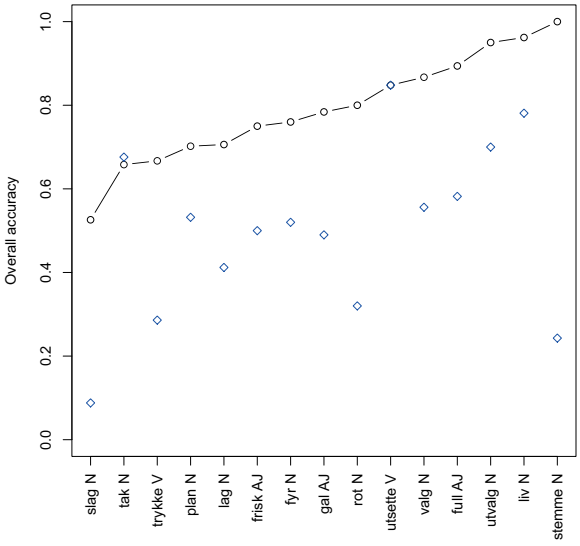
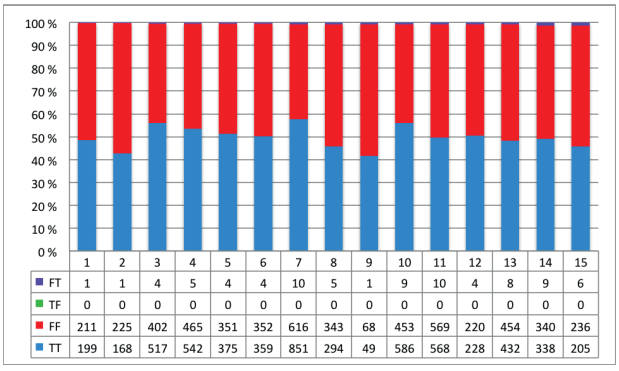



Figure 10.6: Overall Accuracy from Table (10.13) sorted by REL-W. Legend: green=REL-W, blue=UNION.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FT | 1 | 1 | 4 | 5 | 4 | 4 | 10 | 5 | 1 | 9 | 10 | 4 | 8 | 9 | 6 |
| TF | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FF | 211 | 225 | 402 | 465 | 351 | 352 | 616 | 343 | 68 | 453 | 569 | 220 | 454 | 340 | 236 |
| TT | 199 | 168 | 517 | 542 | 375 | 359 | 851 | 294 | 49 | 586 | 568 | 228 | 432 | 338 | 205 |

Figure 10.7: The distribution of known and unknown test lemmas in the best REL-W model per target word vs. its counterpart UNION model.

## 10.5 Discussion and conclusion

The present chapter has presented theoretical experiments aimed at isolating the potential gain in replacing unanalysed context words by Mirrors-derived semantic information about the same words. These are controlled experiments in the sense that instead of collecting context words according to some traditionally used WSD criteria (such as collecting the $n$ nearest open-class words), the presented experiments *only* considered those context words that were known, *a priori*, to have Mirrors-derived sense information in virtue of having been sense-tagged automatically by the tagger in Chapter (7)).

Interestingly, the experiments indicate very much the same pattern as in Chapter (9): the knowledge source SEMANTIC-FEATURES is 'the odd one out' whereas the difference between using semantically unanalysed WORDS and using the RELATED-WORDS is only marginal. Comparing and contrasting the two Mirrors-derived knowledge sources–SEMANTIC-FEATURES (SFs) and RELATED-WORDS (REL-WS)–the results showed that the generalisation gain is low—small groups of context words are grouped in virtue of shared semantic properties according to the Mirrors method, but most words (or rather, most word senses) remain alone in that no other context words during training give support to them. Unfortunately, this is a problem that must be attributed to the corpus size and not to properties of the Mirrors as such, which makes it difficult to make claims as to the generality of the presented observations in the face of a bigger or a different corpus.

Considering SEMANTIC-FEATURES, the results indicate that even though many of the SF-based generalisations are plausible, they mostly seem to have the same unwanted properties that we also find in a traditional W model: some SFs are very frequent (the most frequent SFs have a markedly higher frequency than the most frequent context features in a W model) whereas *most* SFs are extremely low-frequent. Comparing the classification outputs, the W and SF models largely classify on the basis of the same set of known context lemmas. Considering RELATED-WORDS, the results suggest that the gain in adding information from the Mirrors method is extremely marginal. No statistical evidence indicates that there is a difference between a simple WORDS classifier and a RELATED-WORDS classifier. Although not statistically significant, differences are found, but it seems that the amounts of context words that could be grouped were generally on the low side.

# CHAPTER 11

# SUMMARY AND CONCLUDING REMARKS

This study has investigated the use of the translation-based Mirrors method (Dyvik, 2005, *inter alia*) for Word Sense Disambiguation (WSD) for Norwegian. First, the project investigates Dyvik's hypothesis that situated translations constitute an inter-subjective and observable source of information about lexical semantics (the Mirrors method). Second, the Mirrors information about word senses is applied as a knowledge source in a corpus-based machine learning (ML) approach to Word Sense Disambiguation (WSD).

In the following I will first summarise and discuss the findings presented in the preceding chapters Section (11.1). Section (11.2) sums up the principal problems and limitations of the presented thesis. Then some directions for future research will be pointed out (Section (11.3)).

## 11.1 Main contributions and findings

**Overview**

Work on Word Sense Disambiguation (WSD) usually builds on commonly used, and therefore, well-understood sense inventories and data sets. For Norwegian no sense-tagged data previously existed, and there is also no existing WordNet-like resource (although a Norwegian WordNet is expected to be ready in 2012). The presented thesis introduced several new resources in this respect, in that the Mirrors method was explored as an experimental lexical knowledge source, in which WSD was primarily the practical setting for evaluating the use of the Mir-

rors method. As a part of the present thesis a parallel corpus (the ENPC) was automatically sense-tagged (on the English and the Norwegian side) with Mirrors senses, and a lexical sample of fifteen target words was drawn from this material. This formed the basis for a series of controlled experiments, in which the knowledge source to learn from is varied but where we maintain the same experimental framework in terms of the *classification algorithm, data sets, lexical sample and sense inventory*.

**The theoretical motivation for using the Mirrors method in WSD**

Let us first summarise the theoretical motivation for using the Mirrors method in WSD. Some space was devoted to this as the current project is best described as a 'proof of concept experiment' in virtue of introducing several new resources (the Mirrors method, the sense-tagged ENPC and the lexical sample). It was therefore deemed well-motivated to present quite carefully the theoretical foundations and to show why there could, at least in theory, be a gain in abstracting from traditional context words to Mirrors-derived information about the same context words. Due to the 'basic research' nature of the project, furthermore, it is also argued that a lexical sample approach is more suitable than a broad all-words approach to WSD.

Chapter (2) is a reflection on the scientific legitimacy of data-driven language modelling. This was relevant for the present thesis since data-driven methods come to play in two ways in my project: the Mirrors method as well as ML methods for WSD classify data on the basis of evidence from a corpus. The chapter considered the extent to which the current popularity of data-driven methods in modern linguistics represent a revival of 'pre-Chomskyan' ideas, concluding that induction in modern lingustics is not inductive in the traditional, inductivist understanding of the term, due to the difference between *induction* as a formal tool of logical reasoning, as opposed to *inductivism* as a philosophical view on science. It was argued that inductive reasoning as a method does not in itself presuppose that the data sets must be objective (in the inductivist sense of 'objective') in order to generalise from them. It was suggested that induction in modern linguistics is perhaps best seen as a methodological tool for deriving hypothetical generalisations, as an alternative to hypotheses developed through introspection.

Chapter (3) showed how the specific idea of using a structured lexical resource (the Mirrors method) and learning from contextual semantic features is motivated in relation to current state of the art in Word Sense Disambiguation (WSD). The 'standard' approach of applying statistical methods on corpora, typically using WordNet as the main lexical knowledge source, seems to be at a halt in terms of performance. There is therefore an increasing interest in alternative solutions, one of the suggestions being added semantic information. It was also argued that WordNet in many ways offers a kind of structured information that would poten-

tially be very interesting for WSD, but the SENSEVAL competitions have demonstrated that the WordNet senses are too fine-grained for efficient WSD, which has made WordNet less attractive for machine-learnt WSD. Ide and Wilks (2006) recommend instead to approximate word senses by way of cross-lingual sense definitions, which brings us to the chapter about the Mirrors method.

The Mirrors method (Chapter (4)) is developed by Dyvik (1998, *inter alia*) and is an attempt to build knowledge about lexical semantics through translational corpus data. The Mirrors method may be applied for all open-class words (nouns, verbs, adjectives and adverbs) and for any language pair for which parallel corpora exists. It was argued that since adverbs do not seem to generate interesting semantic relations, they were disregarded in the current thesis. The chapter laid out the theoretical foundation for the Mirrors method and considered implementational challenges and solutions in the use of the Mirrors method in the current thesis. Specifically, lemmatisation errors and the automatic word-alignment of the ENPC generates noise in the input material to the Mirrors method, which unfortunately also creates unwanted noise in the Mirrors word bases.

The chapter also showed that in virtue of being a set-theoretic approach, the Mirrors depends on translational overlap and not on statistics: each observation only needs to be recorded once in order to provide useful information. The Mirrors method is thus in principle beneficial in the face of sparse data. Moreover the chapter showed that the output of the Mirrors method resembles a thesaurus or a wordnet entry, containing abstract information about word senses and semantic relations of similarity between word senses. The interesting property of the Mirrors method is that it does not rest on manually derived judgments but is instead based on the *consistent criterion* of translations and translational overlap. It is thus desirable to evaluate how far this criterion takes us.

Chapter (5) first discussed ways of evaluating the Mirrors method, asserting that there are two aspects of the Mirrors method that lend themselves to evaluation, viz. (i) the sense partitions and (ii) the semantic relatedness between senses, that is, the wordnet-like aspect of the Mirrors method. It was pointed out that such a resource is notoriously difficult to evaluate in a fully satisfactory way, partially because there is no commonly agreed upon 'gold standard', and partially because a manual, qualitative evaluation is necessary in order try to separate the implementational circumstances (for instance the modest corpus size, automatic word alignment and lemmatisation errors) from our focus, namely the theoretical Mirrors assumptions.

The chapter then suggests that the monolingual task of WSD could offer a suitable evaluation framework for testing the Mirrors as a knowledge source, since the basic empirical question underlying the Mirrors method is as follows: are the translation-based senses and semantic relations in the Mirrors method linguistically motivated *from a monolingual point of view*? It is shown that the idea that a

well-defined end-user application may provide a stable framework within which the benefits and drawbacks of a resource or a system can be demonstrated also finds support in related work (especially Ng & Lee, 1996; Stevenson & Wilks, 2001; Yarowsky & Florian, 2002; Specia et al., 2009).

**On the experimental framework**

Chapter (6) established a series of controlled experiments to evaluate the Mirrors method, in which the knowledge source to learn from is systematically varied while maintaining the same experimental framework in terms of the classification algorithm, data sets, lexical sample and sense inventory. Three knowledge sources were introduced, one representing a traditional kind of knowledge source that may be derived from a text corpus, and the other two representing Mirrors-derived kinds of knowledge. The three were thus the following:

- Traditional WORD co-occurrences (WS)
  Representing a kind of 'best-known' point of reference to indicate how well a traditional word-based classifier could be expected to perform, given our specific data sample, sense inventory and classification algorithm.

- SEMANTIC-FEATURE (SF)
  An automatically sense-tagged context word is replaced by the Mirrors-derived SFs associated with this word sense.

- RELATED-WORDS(REL-WS)
  The REL-W definition builds on the tentative definitions used in the Mirrors method to discover relations of similarity such as synonymy, hyponymy and hypernymy. The RELATED-WORDS definition, however, neutralises the difference between the various semantic relations being currently explored in the Mirrors method, and instead attempts to select a strict class of semantically related words (regardless if they are hyponyms, hyperonyms or synonyms).

Some examples were also presented to illustrate why one may believe that there could, at least in theory, be a gain in abstracting from traditional context WORDS (WS) to Mirrors-derived information (SEMANTIC-FEATURES or RELATED-WORDS).

The same chapter also motivated the choice to use Naive Bayes as our classification algorithm. Naive Bayes was chosen because of its simplicity. The main emphasis was to identify a well-understood model with good merits in WSD, which is at the same time relatively simple and transparent in terms of analysing the classification model and outcomes. Finally the choice of evaluation metrics was out-

lined and the choice of significance tests was motivated. We will now consider the development of experimental data which is treated in Chapter (7) and Chapter (8).

**Developing data material for WSD experiments for Norwegian**

Chapter (7) and Chapter (8) dealt with the development of a sense-tagged corpus and the selection of a lexical sample to be used in WSD experiments. In the absence of existing data sets for WSD for Norwegian, sense-tagged data and a manually verified lexical sample was developed for Norwegian as part of this thesis. Chapter (7) describes the Mirrors-based automatic sense-tagger, which was applied to the English-Norwegian Parallel Corpus (the ENPC) in order to provide a partially semantically analysed context—partially, because the translation-based sense-tagger can only sense-tag tokens with successful word-alignment. The sense-tagged English-Norwegian Parallel Corpus (the ENPC) is comparable in size to the existing SemCor. As opposed to SemCor, the sense-tagged ENPC has not been manually verified but the thesis demonstrates that it is feasible to produce large corpora on the basis of a word-aligned parallel corpus. The proposed methodology is applicable for any language pair for which word-aligned corpus material exists, and it may then be applied on both language sides. The sense-tagged material constitutes an extension to the existing ENPC in that the sense-tagging results were stored at token level in the XML structure of the ENPC. It has already been suggested in the literature to map sense-tags from one language side to the other side in aligned texts (e.g. Diab & Resnik, 2002; Pianta & Bentivogli, 2003), but the suggested approaches have the possible drawback that the sense-tag of a word in language $L1$ does not necessarily fit as an individual sense for the corresponding $L2$ word. In the proposed approach, each word in each language has a unique set of sense partitions from the Mirrors method which are then used to sense-tag instances.

Chapter (8) advocates that the most sound evaluation setup for this thesis is to aim for a *lexical sample* evaluation. The so-called 'all-words WSD' approach has increasingly come to be the norm because of scalability issues: with a small lexical sample one cannot say whether this approach would scale up. But although it may be legitimate to argue that an *approach* will not scale up (which is especially a challenge for approaches based on manual work), the lexical sample task is then not, by itself, part of the scalability problem. It is furthermore argued that for the present dissertation a lexical sample is well-motivated because the Mirrors method, being an experimental knowledge resource, makes it particularly desirable to focus on a tractable lexical sample to obtain a good analysis of the behaviour of the classifiers.

Since this dissertation has introduced a set of Norwegian target words with a sense inventory that is not commonly known in the WSD community, the target

words and data sets were documented in some detail. The lexical sample contains 15 target words (ten nouns, three adjectives and two verbs). The data set is compiled from the ENPC, consisting of all instances that were sense-tagged automatically and all instances that could not be sense-tagged automatically; the latter were sense-tagged manually. By combining automatically and manually sense-tagged corpus instances we acquire a larger material from the ENPC while maintaining the opportunity to use Mirrors-derived information about context words, since the ENPC is automatically sense-tagged. This is in line with the stated experimental framework of Chapter (6), in which the automatic sense-tagging of Chapter (7) is seen as a separate task from that of systematic experiments with context information. The total data set has on average $269 \pm 337$ corpus instances, the minimum number being 54 instances and the maximum being 1324. It was also pointed out that even though the Mirrors method could enable us to produce large sense-tagged corpora automatically, the usefulness of such tagged data depends on the *plausibility* of the Mirrors method. So the main question, to be pursued in the ensuing chapters, was then: what can we say about the plausibility of the sense-tagged material, and what is the practical applicability of these resources for WSD?

Regarding the lexical sample, it was pointed out that although the size of the data sets are on the low side, a comparison against data sets for other languages indicate that the data sets for Norwegian words are within the limits of what seems to be acceptable in the WSD community. It was also argued that the collection of more material from another resource than the ENPC is beyond the scope of the present project, since the motivating factor for the entire project concerns whether one may 'enlarge' the information value of small data sets by adding Mirrors-derived information about the words surrounding an ambiguous target word. The systematic experiments for testing with and without Mirrors-information about context words makes the current thesis bound to the ENPC, since it is word-aligned (and since it was chosen to experiment with Norwegian material).

**On the Mirrors as a knowledge source in WSD**

Chapter (9) took a traditional WORD classifier as its starting point, testing the effect of replacing the actually occurring lemmas with information from the two Mirrors-derived knowledge sources—SEMANTIC-FEATURES and RELATED-WORDS, when available (EXP1, EXP2 and EXP3, respectively); finally the three knowledge sources were combined in a classification setup where the most confident classifier for each test instance was allowed to vote (EXP4). The findings of this chapter support previously seen evidence, both in (Chapter (7)) and in previous work: approximately half of the lemmas in the ENPC are sense-tagged automatically, which means that quite a lot of information is in principle lost when focussing on Mirrors-specific context information. Seen from this point of view,

it is actually quite encouraging that there is no greater loss in the classification accuracy when abstracting from the nearest words to Mirrors-derived information.

Chapter (10) pruned away those context lemmas that were not sense-tagged, in order to isolate the direct, theoretical effect of replacing the actually occurring words with Mirrors-derived information about the same words (EXP5, EXP6, EXP7). A controlled experiment was also conducted to test the plausibility of the Mirrors word senses (EXP8).

The results from both chapters suggest a similar pattern: the knowledge source SEMANTIC-FEATURES is 'the odd one out' whereas the difference between using semantically unanalysed WORDS and using the RELATED-WORDS is only marginal. Chapter (10) indicated that the generalisation whe using the two Mirrors-derived knowledge sources gain, compared to a simple WORD model, is lower than expected, and the relation between context features and frequencies is, in fact, quite Zipf-like for all three knowledge sources: Some context words are grouped in virtue of shared semantic properties according to the Mirrors method, but most words (or rather, most word senses) remain alone in that no other context words during training give support to them.

As regards the test of the Mirrors sense distinctions in EXP8, the results indicate quite clearly (although the small data sample does not allow strong conclusions) that the best results are given when using sense-specific information, i.e. when trusting the Mirrors senses that are predicted in the context according to the Mirrors-based automatic sense-tagger.

So in response to the question of how well a traditional WORD classifier could be expected to perform, given our specific data sample, sense inventory and classification algorithm, the WORDS classifiers were generally well above the baseline (the most frequent sense baseline). When replacing context words with Mirrors-derived information (EXP2 and EXP3), the experiments did not display strong complementary benefits with respect to the WORDS classifier. On the whole it seems that the three presented knowledge sources are largely in agreement as regards what they learn, but the SFs seem to sometimes suffer from using too general SFs.

For the sake of WSD and the hope to improve the results by adding paradigmatic knowledge, then, the missing *gain* may appear disappointing. But as far as the plausibility of the Mirrors method is concerned, this finding is quite promising, since this means that even when some context words are lost, the RELATED-WORDS knowledge source still seems to obtain the same, necessary information as a traditional WORDS model. The missing difference means that there were no findings to indicate serious drawbacks of the principles underlying the Mirrors method. Although the amounts of related word senses are quite modest, the results also demonstrated that the Mirrors method does succeed in deriving relations of similarity between words that are also discovered using corpus-based principles of co-occurrences.

Some of the classification examples did *not* seem to support the assumption that if a context word $x$ co-occurred with a target word, then so can also probably the semantic relatives of $x$. A further study of this assumption could be among the factors to be pursued in future work, for instance by testing the same assumption using the Princeton WordNet and SemCor (Section (11.3).

## 11.2 Thesis problems and limitations

- The work has shown that poor quality input to the Mirrors is unfortunate, since the method is vulnerable to noise (cf. problems with the lemmatiser and the automatic word-aligner). It remains unasserted how much such errors influence the Mirrors method, as there was no straightforward way of asserting formally whether suprising semantic relations find their explanation in errors in the Mirrors input or if there are problems with the theoretical Mirrors assumptions.

- The current use of the Mirrors method, and therefore also the presented automatic sense-tagger, is limited by its dependence on available parallel text.

- The quality of the developed lexical sample for Norwegian is evaluated with only one annotator, therefore an inter-annotator agreement measure is not available.

- Although it is in principle an advantage that the Mirrors method can derive plausible information from even small data samples, being independent of statistics, it it not clear how the Mirrors method would perform with significantly larger data material than the presented use of the ENPC. Testing on an independent, larger sample might shed light on this.

- Since the sense-tagger and the ensuing WSD experiments are tested on a relatively small corpus resource, it turned out that we could not draw clear conclusions regarding the true quality of the Mirrors assumptions.

## 11.3 Future work

Viewing the presented work as a 'proof of concept' experiment regarding the use of the Mirrors method for WSD, I leave as questions for future research the following:

- Testing the Mirrors method and the automatic sense-tagger on different language pairs and on different corpus data:
  An interesting further path would be to test, first, the Mirrors method on different parallel corpus data. The behaviour of the Mirrors method on other language pairs and with different corpus samples is not well-understood to date. For instance, the Europarl material from SEMEVAL-2010 could provide some interesting material for future work.

- Testing the presented classifiers on comparable corpora:
  The time frame of the current thesis did not permit testing the classifiers on an independent, comparable corpus. A comparable corpus is typically defined as belonging to the same genre and having approximately the same size. Monolingual corpora for Norwegian exists that may be defined as comparable to the Norwegian side of the ENPC, but they are not sense-tagged.

- Testing the presented classifiers for bootstrapping on comparable corpora:
  By applying the classifiers of this thesis as bootstrapping seeds, one may use bootstrapping techniques on a comparable corpus to collect more context data.

- Explore the combinations of knowledge sources:
  All the presented classifiers use open-class words only. The classification analysis has shown that for the sake of WSD performance in itself (and not for the sake of evaluating the Mirrors method) the overall accuracies would probably have been improved if local collocations had been included as a source of knowledge.

- Experiment with feature selection:
  The presented experiments specifically chose not to prune away uninformative context features *a priori* since it was a point to be able to consider the direct effect when moving from WORDS to Mirrors-derived information. For future experiments, an *a priori* feature selection would clearly be desirable. It could, in particular, give some new insights concerning the Mirrors method, since feature selections would provide an easy way to visualise whether Mirrors-derived context features more often confuse the target word senses compared to the actually occurring WORDS (information measures are designed to select those features that are particularly characteristic of a specific target word sense). The experiments also showed that the contribution of individual context features does not in itself depend on higher frequencies but on the ratio between the Maximum Likelihood Estimation (MLEs): the higher the ratio, the more it 'pulls' in the direction of

a given sense of the target word. Hence, when using Naive Bayes this property could be exploited during feature selection by sorting first and foremost according to the 'pulling strength'.

- Testing added semantic information using WordNet and SemCor:
  Since added semantic information is one of the suggestions often being proposed in response to the problem that corpus data do not seem to suffice within ML approaches to WSD, it would be interesting to study the abstraction gain from word co-occurrences to classes of semantically related words using a well-known lexical resource (the Princeton WordNet) and a manually verified sense-tagged corpus (SemCor). Such a study might shed further light on the extent to which it *is* well-motivated to exploit paradigmatic relations in WSD: to what extent does it occur that two words may be plausibly semantically related in the paradigmatic, without equally plausibly sharing co-occurrence properties? Such a study might provide valuable insights both for WSD concretely and, more broadly, for lexical semantics studies.

- Explore the discovery of multiword expressions prior to WSD:
  Although this has not been an explicit topic of this thesis, a recurring observation has been that it is not unproblematic to approach lexical meaning only at the level of singular words, which has motivated an ensuing interest in the phenomenon of multiword expressions (MWEs) (Lyse & Andersen, forthcoming). There is in general an emerging awareness that multiword expressions, such as fixed expressions, are not just sporadic exceptions in our vocabulary (e.g. Biber, 2009; Tognini-Bonelli, 2001; Sinclair, 1996; Stubbs, 1996). MWEs are surprisingly ubiquitous in natural language, being estimated to be as frequent as one-word expressions (Jackendoff, 1997); similarly Sag et al. (2002) assert that 41 per cent of the entries in WordNet (Fellbaum, 1998) are multiword units. With respect to the Mirrors, the theoretical assumptions concern translational correspondences between linguistic *signs*, i.e. 'meaning units'. MWEs were therefore included under certain conditions when extracting translational correspondents manually (e.g. Dyvik, 2005; Lyse, 2003). However, MWEs were beyond the scope of the automatic word-alignment: As was seen in Chapter (8), an analysis of the instances that could not be sense-tagged automatically (because they were not word-aligned) showed that fixed expressions were found in the *untagged* test sets, and not in the automatically sense-tagged sets. In other words, fixed expressions seem to be a problem for automatic word alignment. An interesting pre-processing step could therefore be to attempt to discover larger lexical units (MWEs), and then to run automatic word-alignment.

# APPENDIX A

# APPENDICES

All appendices are downloadable in PDF from
URL: http://hdl.handle.net/1956/4712

# REFERENCES


Abney, S. (2000). *Statistical methods*. Nature Publishing Group, Macmillan.

Agirre, E. & Edmonds, P. (2006a). Introduction. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 1). Springer.

Agirre, E. & Edmonds, P. (Eds.). (2006b). *Word Sense Disambiguation: Algorithms and applications*. Springer.

Agirre, E., Magnini, B., Lacalle, O. Lopez de, Otegi, A., Rigau, G. & Vossen, P. (2007, June). SemEval-2007 task 01: Evaluating WSD on cross-language information retrieval. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 1–6). Prague, Czech Republic: Association for Computational Linguistics.

Agirre, E. & Martínez, D. (2000). Exploring automatic Word Sense Disambiguation with decision lists and the Web. In *Proceedings of the semantic annotation and intelligent annotation workshop, organized by coling* (p. 11-19). Luxembourg.

Agirre, E. & Martínez, D. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 25–32). Barcelona, Spain.

Agirre, E. & Rigau, G. (1996). Word Sense Disambiguation using conceptual density. In *Proceedings of the 16th conference on computational linguistics* (pp. 16–22). Morristown, NJ, USA: Association for Computational Linguistics.

Agirre, E. & Soroa, A. (2007, June). SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 7–

12). Prague, Czech Republic: Association for Computational Linguistics.

Agirre, E. & Stevenson, M. (2006). Knowledge sources for WSD. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 8). Springer.

Apidianaki, M. (2008, may). Translation-oriented word sense induction based on parallel corpora. In N. C. (Chair) & colleagues (Eds.), *Proceedings of the sixth international language resources and evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA). (http://www.LREC-conf.org/proceedings/LREC2008/)

Bakx, G. E., Villodre, L. M. & Claramunt, G. R. (2006). *Machine learning techniques for Word Sense Disambiguation*. Unpublished doctoral dissertation, Universitat Politecnica de Catalunya.

Banerjee, S. & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on artificial intelligence (ijcai,* (pp. 805–810). Acapulco, Mexico.

Bar-Hillel. (1960). Automatic translation of languages. In D. Booth & R. E. Meagher (Eds.), *Advances in computers.* Academic.

Biber, D. (2009). A corpus-driven approach to formulaic language in english: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311.

Bloomfield, L. (1933). *Language*. London: Ruskin House; George Allen and Unwin Ltd.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. & Mercer, R. (1991). Word Sense Disambiguation using statistical methods. In *Proceedings of the 32. annual meeting of the Association for Computational Linguistics.* Las Crues, New Mexico.

Bruce, R. & Wiebe, J. (1994). Word Sense Disambiguation using decomposable models. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics* (pp. 139–146). Las Cruces, New Mexico, USA.

Chan, Y. S. & Ng, H. T. (2005). Scaling up Word Sense Disambiguation via parallel texts. In *Aaai'05: Proceedings of the 20th national conference on artificial intelligence* (pp. 1037–1042). AAAI Press.

Charles Fillmore, C. R. J. & Petruck, M. R. (2003). Background to Framenet. *International Journal of Lexicography*, *16*, 235 – 250.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MA: The MIT Press.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. New York and London: Harper and Row.

Cucchiarelli, A. & Velardi, P. (2002). Feature-based WSD: Why we are at a dead-end. In *Portal '02: Proceedings of the third international conference*

*on advances in Natural Language Processing* (pp. 5–14). London, UK: Springer-Verlag.

Daelemans, W. & Bosch, A. Van den. (2005). *Memory-based language processing*. New York, NY, USA: Cambridge University Press.

Daelemans, W., Bosch, A. van den & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. In *Machine learning, special issue on natural language learning* (Vol. 34, pp. 11–41). Hingham, MA, USA: Kluwer Academic Publishers.

Daelemans, W., Zavrel, J., Sloot, K. V. der & Bosch, A. V. den. (2007). *TiMBL: Tilburg memory based learner, version 6.1, reference guide* (ILK Technical Report No. 07-07). Tilburg.

Dagan, I. (1991). Lexical disambiguation: sources of information and their statistical realization. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics.* Berkeley, California.

Dagan, I. & Itai, A. (1994). Word Sense Disambiguation using a second language monolingual corpus.

Diab, M. & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In J. Carroll, N. Oostdijk & R. Sutcliffe (Eds.), *Acl 2002, proceedings of the 40th annual meeting of the Association for Computational Linguistics.* Philadelphia, Pennsylvania.

Dyvik, H. (1997). *Data, facts and concepts of language.* (Lecture at the summer school on Language Data and Linguistic Questions, Bergen, 1997)

Dyvik, H. (1998). A translational basis for semantics. In S. Johansson & S. Oksefjell (Eds.), *Corpora and crosslinguistic research: Theory, method and case studies* (pp. 51 – 86). Rodopi.

Dyvik, H. (2004). Translations as semantic mirrors. From parallel corpus to WordNet. *Language and Computers*, *1*, 311 – 326.

Dyvik, H. (2005). Translations as a semantic knowledge source. In *Proceedings of the second baltic conference on human language technologies.* Tallinn University of Technology.

Dyvik, H. (2009). *Semantic mirrors.* ([Draft])

Fellbaum, C. (Ed.). (1998). *WordNet. An electronical lexical database*. MIT Press.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis, Philological Society, Oxford. In E. by F. R. Palmer 1968 (Ed.), *(reprinted in 1968 in selected papers of j. r. firth, 1952-59).* Indiana University Press (Bloomington).

Gale, W. A., Church, K. W. & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, *26*, 415 – 439.

Glizzio, A., Giuliano, C. & Strappavara, C. (2005). Domain kernels for Word

Sense Disambiguation. In *Proceedings of the 43rd annual meeting of the acl* (pp. 403—410). Association for Computational Linguistics.

Harris, R. (1993). *The linguistics wars*. New York: Oxford University Press.

Harris, Z. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hearst, M. (1998). Automated discovery of WordNet relations. In C. Fellbaum (Ed.), *WordNet. an electronical lexical database* (pp. 131 – 152). MIT Press.

Hofland, K. & Johansson, S. (1998). The translation corpus aligner: A program for automatic alignment of parallel texts. In S. Johansson & S. Oksefjell (Eds.), *Corpora and crosslinguistic research: Theory, method, and case studies* (p. 87-100). Amsterdam: Rodopi.

Hoste, V., Hendrickx, I., Daelemans, W. & Bosch, A. van den. (2002). Parameter optimization for machine-learning of Word Sense Disambiguation. *Natural Language Engineering*, *8*(4), 311-325.

Hoste, V., Hendrickx, I., Daelemans, W. & Van Den Bosch, A. (2002). Parameter optimization for machine-learning of Word Sense Disambiguation. *Natural Language Engineering*, *8*(4), 311–325.

Ide, N. (1999). Parallel translations as sense discriminators. In *Proceedings of the acl siglex workshop on standardizing lexical resources* (pp. 52 – 61). College Park, Maryland.

Ide, N. & Erjavec, T. (2001). Automatic sense tagging using parallel corpora. In *Proceedings of the 6 th Natural Language Processing pacific rim symposium* (pp. 83–89). The MIT Press.

Ide, N., Erjavec, T. & Tufiş, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of acl-02 workshop on Word Sense Disambiguation: Recent successes and future directions* (pp. 54–60). Philadelphia.

Ide, N. & Veronis, J. (1990, October). Mapping dictionaries: A spreading activation approach. In *Proceedings of the 6th conf. for the new oed conference* (pp. 52–64). Waterloo.

Ide, N. & Véronis, J. (1998). Introduction to the special issue on Word Sense Disambiguation: the state of the art. *Computational Linguistics*, *24*(1), 2–40.

Ide, N. & Wilks, Y. (2006). Making sense about sense. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 3). Springer.

Izquierdo, R., Suárez, A. & Rigau, G. (2007, June). GPLSI: Word coarse-grained disambiguation aided by basic level concepts. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 157–160). Prague, Czech Republic: Association for Computational Linguistics.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge,

MA: The MIT Press.

Johansson, S., Ebeling, J. & Oksefjell, S. (1999/2002). English-Norwegian Parallel Corpus: Manual [Computer software manual]. Oslo, Norway.

Kaplan, A. (1950). *An experimental study of ambiguity in context.* cited in Mechanical Translation, v. 1, nos. 1-3.

Kilgariff, A. (2006). Word senses. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 2). Springer.

Klein, D. E. & Murphy, G. L. (2002). *Paper has been my ruin: Conceptual relations of polysemous senses* (Vol. 47).

Lager, T. & Zinovjeva, N. (2001, July 5-6, 2001). Sense and deduction: The power of peewees applied to the SENSEVAL-2 Swedish lexical sample task. In *Proceedings of SENSEVAL-2: Second international workshop on evaluating Word Sense Disambiguation systems.* Toulouse, France.

Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for Word Sense Disambiguation. In C. Fellbaum (Ed.), *WordNet. an electronical lexical database* (pp. 265 – 283). MIT Press.

Leacock, C., Miller, G. A. & Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, *24*(1), 147–165.

Leacock, C., Towell, G. & Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Hlt '93: Proceedings of the workshop on human language technology* (pp. 260–265). Morristown, NJ, USA: Association for Computational Linguistics.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M. et al. (2000, December). SIMPLE: A general framework for the development of multilingual lexicons. *Int J Lexicography*, *13*(4), 249–263.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Sigdoc '86: Proceedings of the 5th annual international conference on systems documentation* (pp. 24–26). New York, NY, USA: ACM.

Lyse, G. I. (2003). *Fra speilmetoden til automatisk ekstrahering av et betydningstagget korpus for WSD-formaal. [from the mirrors method to automatic extraction of sense-tagged corpora for WSD].* Unpublished master's thesis, Linguistics department, University of Bergen, Norway.

Lyse, G. I. (2006, July). "Making Sense of Translations"–Translation-based lexical information for Word Sense Disambiguation (WSD). In E. T. Vold, G. I. Lyse & A. M. Gjesdal (Eds.), *New voices in linguistics* (pp. 233 – 246). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Lyse, G. I. & Andersen, G. (forthcoming). Collocations and statistical analysis of n-grams. In G. Andersen (Ed.), *Exploring newspaper language - corpus compilation and research based on the norwegian newspaper corpus.* John

Benjamins.

Magnini, B., Strapparava, C., Pezzulo, G. & Gliozzo, A. (2002). The role of domain information in Word Sense Disambiguation. *Nat. Lang. Eng.*, *8*(4), 359–373.

Manning, C. & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge: Massachusetts: MIT Press.

Márquez, L., Escudero, G., Martinez, D. & Rigau, G. (2006, July). Supervised corpus-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (pp. 167–207). Springer.

Masterson, M. (1967). Mechanical Pidgin translation. In D. Booth (Ed.), *Machine translation.* Wiley, New York.

Mihalcea, R. (2002a). Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd international conference on language resources and evaluations (LREC).* Las Palmas.

Mihalcea, R. (2002b). Word Sense Disambiguation with pattern learning and automatic feature selection. *Nat. Lang. Eng.*, *8*(4), 343–358.

Mihalcea, R. (2006, July). Knowledge-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (pp. 107–131). Springer.

Mihalcea, R. & Faruque, E. (2004, July). SenseLearner: Minimally supervised Word Sense Disambiguation for all words in open text. In R. Mihalcea & P. Edmonds (Eds.), *SENSEVAL-3: Third international workshop on the evaluation of systems for the semantic analysis of text* (pp. 155–158). Barcelona, Spain: Association for Computational Linguistics.

Mihalcea, R. & Moldovan, D. (1999). An automatic method for generating sense tagged corpora. In *Proceedings of aaai-99* (pp. 461–466). AAAI Press.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet. an electronical lexical database* (chap. 1). MIT Press.

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

Montoyo, A. & Suárez, A. (2001). The University of Alicante Word Sense Disambiguation system. In J. Preiss & D. Yarowsky (Eds.), *Proceedings of SENSEVAL-2, acl-siglex* (pp. 131–134). Toulouse, France.

Montoyo, A., Suarez, A. & Rigau, G. (2005). Combining knowledge- and corpus-based Word Sense Disambiguation methods. *Journal of Artificial Intelligence Research*, *23*, 2005.

Mooney, R. J. (1996, May). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the conference on empirical methods in Natural Language Processing* (pp. 82–91). Philadelphia, PA.

Navigli, R. (2009). Word Sense Disambiguation: A survey. *ACM Comput. Surv.*, *41*(2), 1–69.

Ng, H. T. (1997a). Exemplar-based Word Sense Disambiguation: Some recent improvements. In C. Cardie & R. Weischedel (Eds.), *Second conference on empirical methods in Natural Language Processing (EMNLP –2)* (pp. 208 – 213). Somerset, New Jersey: Association for Computational Linguistics.

Ng, H. T. (1997b, April). Getting serious about Word Sense Disambiguation. In *Association for Computational Linguistics special interest group on the lexicon (acl-siglex-1997): Workshop"tagging text with lexical semantics: Why, what, and how?"* (pp. 1–7). Washington, D.C., USA.

Ng, H. T. & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting of the assocation for computational linguistics* (pp. 40–47). Santa Cruz, California, USA: Morgan Kaufmann Publishers Inc.

Nivre, J. (2002). On statistical methods in Natural Language Processing. In J. jr Bubenko & B. Wangler (Eds.), *Promote IT. Second conference for the promotion of research in IT at new universities and university colleges in Sweden, University of Skoevde* (pp. 694 – 694). Sweden.

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosén, V. & Flickinger, D. (2007). Towards hybrid quality-oriented machine translation — on linguistics and probabilities in MT. In *Proceedings of the 11th conference on theoretical and methodological issues in machine translation (TMI-07).* Skövde, Sweden.

Palmer, M., Ng, H. T. & Dang, H. T. (2006). Evaluation of WSD systems. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 4). Springer.

Patwardhan, S., Banerjee, S. & Pedersen, T. (2005). SenseRelate::TargetWord: A generalized framework for Word Sense Disambiguation. In *Acl '05: Proceedings of the acl 2005 on interactive poster and demonstration sessions* (pp. 73–76). Morristown, NJ, USA: Association for Computational Linguistics.

Patwardhan, S., Banerjee, S. & Pedersen, T. (2007, June). UMND1: Unsupervised Word Sense Disambiguation using contextual semantic relatedness. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 390–393). Prague, Czech Republic: Association for Computational Linguistics.

Pedersen, T. (1999). *Search techniques for learning probabilistic models of Word Sense Disambiguation.* (Working Notes of the AAAI Spring Symposium on Search Techniques for Problem Solving Under Uncertainty and Incomplete Information, March 22-24, Palo Alto, CA.)

Pedersen, T. (2000). A simple approach to building ensembles of naive Bayesian

classifiers for Word Sense Disambiguation. In *Proceedings of the first conference of the north american chapter of the Association for Computational Linguistics* (pp. 63–69). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pedersen, T. (2002). Assessing system agreement and instance difficulty in the lexical sample tasks of SENSEVAL-2. In *Proceedings of the acl-02 workshop on Word Sense Disambiguation* (pp. 40–46). Morristown, NJ, USA: Association for Computational Linguistics.

Pedersen, T. (2006). Unsupervised corpus-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 6). Springer.

Pianta, E. & Bentivogli, L. (2003). Translation as annotation. In *Proceedings of the aiia 2003, workshop "topics and perspectives of Natural Language Processing"*. Pisa, Italy.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson & Co.

Priss, U. & Old, L. J. (2005). Conceptual exploration of Semantic Mirrors. In *Formal concept analysis: Third international conference, icfca 2005.* Springer Verlag.

Resnik, P. (1995). Disambiguating noun groupings with respect to WordNet senses. In D. Yarovsky & K. Church (Eds.), *Proceedings of the third workshop on very large corpora* (pp. 54 – 68). Somerset, New Jersey: Association for Computational Linguistics.

Resnik, P. (2004). Exploiting hidden meanings: Using bilingual text for monolingual annotation. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing (cicling)* (pp. 283–299). Springer-Verlag.

Resnik, P. (2006). WSD in NLP applications. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and applications* (chap. 11). Springer.

Resnik, P. & Yarowsky, D. (1997, 1999). Distinguishing systems and distinguishing senses: New evaluation methods for Word Sense Disambiguation. *Nat. Lang. Eng.*, *5*(2), 113–133.

Rowntree, D. (2000). *Statistics without tears: An introduction for non-mathematicians*. Penguin Books.

Sæbø, K. J. (2004). Natural language corpus semantics: The free choice controversy. *Nordic Journal of Linguistics*, *27*(02), 197–218.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the third international conference on intelligent text processing and computational linguistics.* Mexico City, Mexico.

Sinclair, J. M. (1996). The search for units of meaning. In *Textus* (Vol. 9, pp. 75–106).

Specia, L., Graças, M., Nunes, V. & Stevenson, M. (2008). *A hybrid relational approach for Word Sense Disambiguation.* Available from www.xrce.xerox.com/content/download/7021/52505/file/2008-045.pdf

Specia, L., Graças, M. D., Nunes, V. & Stevenson, M. (2005). Exploiting parallel texts to produce a multilingual sense tagged corpus for Word Sense Disambiguation. In *Proceedings of raNLP -05, borovets* (pp. 525–531). Borovets.

Specia, L., Stevenson, M. & Das Graças Volpe Nunes, M. (2009). Assessing the contribution of shallow and deep knowledge sources for Word Sense Disambiguation. In *Language resources and evaluation* (Vol. 44, p. 295-313). Springer (SpringerLink). Available from http://www.springerlink.com/content/00183410568p3553/

Stevenson, M. (2003). *Word Sense Disambiguation — the case for combinations of knowledge sources*. CSLI Publications.

Stevenson, M. & Wilks, Y. (2001). The interaction of knowledge sources in Word Sense Disambiguation. *Computational Linguistics*, *27*(3), 321–349.

Stubbs, M. (1996). *Text and corpus analysis*. London: Blackwell.

Thunes, M. (2003). *Evaluating thesaurus entries derived from translational features.* Paper presented at NoDaLiDa 2003, Reykjavik.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tufiş, D., Ion, R. & Ide, N. (2004). Fine-grained Word Sense Disambiguation based on parallel corpora, word alignment, word clustering and aligned WordNets. In *Proceedings of the 20 th international conference on computational linguistics, coling 2004* (pp. 1312–1318). Morristown, NJ, USA: Association for Computational Linguistics.

Villarejo, L., Màrquez, L. & Rigau, G. (2005, September). Exploring the construction of semantic class classifiers for WSD. In *Procesamiento del lenguaje natural* (pp. 195–202). Granada, Spain.

Vossen, P. (Ed.). (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Norwell, MA, USA: Kluwer Academic Publishers.

Wang, X. & Carroll, J. (2005). Word Sense Disambiguation using sense examples automatically acquired from a second language. In *Hlt '05: Proceedings of the conference on human language technology and empirical methods in Natural Language Processing* (pp. 547–554). Morristown, NJ, USA: Association for Computational Linguistics.

Weaver, W. (1955). Translation. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: fourteen essays* (pp. 15–23). New York: John Wiley & Sons. (Reprint of a text first printed 15 July 1949)

White, J. S. (1988). Determination of lexical semantic relations for multi-lingual

terminology structures. , 183–198.

Wilks, Y. & Stevenson, M. (1996). *The grammar of sense: Is word sense tagging much more than part-of-speech tagging?* (Technical Report No. CS-96-05). University of Sheffield.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th conference on computational linguistics* (pp. 454–460). Morristown, NJ, USA: Association for Computational Linguistics.

Yarowsky, D. (1993). One sense per collocation. In *Hlt '93: Proceedings of the workshop on human language technology* (pp. 266–271). Morristown, NJ, USA: Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting of the Association for Computational Linguistics* (pp. 189–196). Morristown, NJ, USA.

Yarowsky, D. & Florian, R. (2002, December). Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, *8*, 293–310. Available from `http://portal.acm.org/citation.cfm?id=973775.973778`

Yngve, V. (1955). Syntax and the problem of multiple meaning. In W. N. Locke & D. Booth (Eds.), *Machine translation of languages.* Wiley, New York.

Zaanen, T. G. van. (2004). *Linguistic knowledge and Word Sense Disambiguation*. Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands, The Netherlands.

Zhong, Z. & Ng, H. T. (2009). Word Sense Disambiguation for all words without hard labor. In *Ijcai'09: Proceedings of the 21st international jont conference on artifical intelligence* (pp. 1616–1621). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Zipf, G. K. (1935). *The psychobiology of language*. Houghton-Mifflin.