Chapter 2

Lexico-semantic knowledge

2.1 Introduction

In the previous chapter we explained that we want to automatically acquire lexico-semantic knowledge. We explained that lexico-semantic knowledge comprises information about semantic relations between lexical elements. In this chapter we will give some background information on lexico-semantic knowledge. We will introduce the different lexical elements and the semantic relations between them that will be studied in this thesis.

An example of a resource of lexico-semantic knowledge that the reader might be familiar with is Princeton WordNet (Fellbaum, 1998). Princeton WordNet is an electronic resource inspired by current psycholinguistic theories of human lexical memory. Synonyms are grouped in synsets, i.e. lists of words that are (near)-synonyms. These synsets are in turn related by basic semantic relations. We can, for example, find that a cat is a carnivore because there is a semantic relation, i.e. the HYPERNYM RELATION, between carnivore and cat. The hypernym relation puts the word cat in the category of carnivores.

The metaphor of a graph helps us to talk about the lexical elements and the relations between them. The lexical elements are the nodes and the relations between them are the arcs connecting the nodes. We have sketched an example graph in Figure 2.1. After having explained the arcs and nodes, i.e. the lexical elements (section 2.2) and the semantic relations between them (section 2.3), we will give an overview of existing resources in section 2.4. In section 2.5 we will conclude by discussing possible ways of evaluating the acquired lexico-semantic knowledge.
Chapter 2. Lexico-semantic knowledge

2.2 Lexical elements

Before discussing the different kind of relations that exist between lexical elements, we need to define what lexical elements will be the focus of the current study. What is the nature of the elements that we expect to find relations between? We will be concerned with open-class words and we will lemmatise the words included. We will explain these terms briefly in section 2.2.1 and we will discuss the problem that lexical ambiguity poses in section 2.2.2.

2.2.1 Open-class words

In this study we are concerned with open-class words. Examples are nouns, such as *bier* ‘beer’, adjectives, such as *sterk* ‘strong’, and verbs, such as *zien* ‘see’. They belong to the class of words that is open to new words. Van Dale published a top ten of the most widely used new terms in the media for 2007. The verb *sonjabakkeren* is at position number five. *Sonjabakkeren* refers to a special form of dieting introduced by Sonja Bakker.

Open-class words are opposed to closed-class words, such as the determiners *de* ‘the’, and *een* ‘a’, and the conjunction *en* ‘and’. Closed class words are typically more frequent, but there are fewer of them. Acquiring the semantics of the closed class words is something that could in principle be done by hand. For a large class, such as the open-class words, this is less feasible.

In this thesis we will focus on finding semantic relations such as synonymy between open class words and we will dedicate most time to nouns, such as *cat*, and proper names, such as *Groningen*.

The term open-class words is not precise enough. Both *obeys* and *obey* are words. We will incorporate only the canonical form of words in the knowledge base and not all inflected or derived forms. Such canonical forms are often referred to by the term lemma or lexeme. We have chosen the singular form in the case of nouns and the first person singular, present tense as the basic lexical element in the case of verbs. In the example given above we would select *obey* as the lexical element.

The *s* in *obeys* is an inflectional affix. We abstract away from such inflectional affixes by including only lemmas in the knowledge base. After all, we are interested in the meaning of words and inflection has little effect on the
meaning of a word, only on certain aspects, such as tense. On the contrary, derivational affixes distinguish between the meaning of words. They distinguish between syntactic categories as well. Consider the stem help combined with derivational affixes -ful and -er. The word helpful is an adjective, whereas helper is a noun. Furthermore, the common aspect of these words, that is linked to the stem help, is only a part of their full meaning. This explains also why these derivational variations are usually listed as separate items in the dictionary and not as variations of the lexical element help. We will also list them as separate entries.

The verb help brings us to another important feature of the knowledge base. We disambiguate words with respect to the syntactic category they are associated with, if this information is available. For example, the word help can both refer to the verb and the noun reading. There will be separate entries for the verb help and the noun help, if this information is available.

Apart from single words we have also included some multiword terms. However, we limited ourselves to the inclusion of multiword terms that our dependency parser recognises, for example, proper names such as Michael Jackson or Den Haag ‘The Hague’. Although we will refer to the lexical elements in the knowledge base as words in the next sections, it should be clear to the reader that we do not only include single words, but also multiword terms.

2.2.2 Polysemy and homonymy

‘One of the basic problems of lexical semantics is the apparent multiplicity of semantic uses of a single word form (…)’ Cruse (1986). These semantic uses are generally referred to by the term senses. An example of a word with multiple senses is the word bank. The word can either refer to a shore of a river or an establishment for the custody of money.

A distinction is often made between two forms of lexical ambiguity: polysemy and homonymy. In the case of polysemy the several meanings are related, whereas in the case of homonymy they are not. We have already introduced the lemma or lexeme, the canonical form of a set of word-forms, that is used in dictionaries. In case a single lexeme has many senses we speak of polysemy. If a word-form belongs to more than one lexeme we speak of homonymy. However, the ‘border-line (…) is sometimes fluid.’ (Ullman, 1957). In this work we do not make the distinction between polysemy and homonymy. We will speak of polysemy, referring to both related multiple meanings and unrelated multiple meanings.

It would be ideal if we could have a disambiguated account of words and the relations between them. For example, by having an entry for each sense of each
Chapter 2. Lexico-semantic knowledge

word in our knowledge base: bank[1] for the shore of a river, and bank[2] for the custody of money. This would result in having distinct nodes and arcs departing from these nodes for each lexical element. It is what a hand-built lexical resource, such as WordNet (Fellbaum, 1998), tries to do and we will come to speak about it later. This disambiguated account of words, however, requires word sense discovery, which is a study in itself. It falls outside the scope of this thesis.\footnote{The knowledge base to be developed here makes no sense distinctions. We will come across deficiencies resulting from polysemy in the next chapters.}

It might be a poor consolation to note that it is often difficult to make use of the sense distinctions comprised in the knowledge base, when used in an application. Distinguishing between senses of words, when building a knowledge base, is one thing, but making use of this sense information is another thing. Making use of sense information in resources in an adequate way requires word sense disambiguation. Yet another field of study that falls outside the scope of this chapter.

2.3 Lexico-semantic relations

Now that we have made clear what lexical elements we will consider in this thesis, we will discuss the relations between these lexical elements. These relations determine the structure of the lexico-semantic knowledge base. To continue the metaphor of a word graph we introduced in the previous section: We explained what the nodes in our graph are and will now talk about the arcs that connect the nodes.

There are several types of lexico-semantic relations. Kilgarriff and Yallop (2000) use the terms loose and tight to describe different types of lexico-semantic resources. In a loose resource, such as the Roget Thesaurus (Roget, 1911), words are related in an associative way. They are related according to subject field, whereas tight resources tend to group words that are the same kind of things, i.e. that belong to the same semantic class, together.

We include words at increasing levels of tightness in the lexico-semantic knowledge base. In our discussion of the different types of lexico-semantic relations we will go from loose relations (the associative relation, section 2.3.1) to tighter relations (taxonomically related words, section 2.3.2) to an even tighter relation (synonymy, section 2.3.3).
2.3. Lexico-semantic relations

### 2.3.1 Associative relations

Some words are related in an associative way, for example *hospital* and *nurse*. The same holds for *food* and *hunger*. The words do not have to belong to the same semantic class. *Food* belongs to the class of concrete objects, whereas *hunger* is something abstract. A nurse is a human being, whereas a hospital is a building. They are, however, related with respect to subject. When people are in a conversation about hospitals, it is likely that they will speak about nurses and diseases as well. It is less probable that they will start talking about parsley and cooking utensils without introducing a change in subject.

Psychologists have designed free association tests to elicit these associations from human subjects. Participants are asked to respond to a stimulus word with the words that the stimulus word evokes in their mind. We will discuss a free association test for Dutch in section 2.4.2.

We talked about nodes and arcs, the nodes being the lexical elements and the arcs being the lexico-semantic relations. In the case of associative relations we can think of a graph of lexical elements that are at a certain distance to each other depending on the strength of association between them. In figure 2.2 a fragment of this graph is depicted. *Hospital* is closely related to *doctor, nurse* and *disease*, and much less related to *saucer* and *parsley*.

Note that the graph is an undirected graph. Although we are aware of the fact that the human brain does not work in the same way, the system we introduce below produces symmetrical associative relations. According to our system, *nurse* is as much related to *hospital* as *hospital* is to *nurse*.

### 2.3.2 Taxonomically related words

Whereas the associative relations are represented as a flat network of lexical elements at a certain distance, taxonomical relations give rise to a hierarchical structure. Here it is not the subject-relatedness that brings the words together, but the fact that they belong to the same semantic class. Because some classes

---

1. However, in the last chapter we will show some preliminary results.
are supersets of other classes, a hierarchy is born.

We will try to make the distinction between the taxonomic and the associative relation clearer with some discussion. If we look at figure 2.3, we see that within the group of animals we can distinguish the vertebrates and the invertebrates. Within the group of vertebrates we will find the fish, and the mammals. Within the group of mammals we find the carnivores and insectivores. The image of an upside-down tree with branches expanding to the bottom helps to understand the nature of these taxonomic relations. At each point where two branches meet we find the nodes. At the very end of the branches we find nodes as well, but these do not expand to any other branches and are called the leaves. In the example one of the leaves under dog is Fluffy a name that refers to an instance of dog, a particular dog that exists in the world at some point in time.

Let us again introduce some terminology. If two nodes are connected by one single branch the more general node is called the mother node and the more specific node the child node. More general nodes that can be reached from a certain child node without having to change direction (going from more specific to more general) are a node’s ancestors. On the other hand, all nodes under a certain mother node that can be reached without having to change direction are called its descendants. These terms invoke the analogy with a family and its members. However, in a family descendants follow from a mother and a father.
In the example in 2.3 each node descends from one single lexical element. The best analogy would be that of a single parent family, such as a family of starfish. Certain paths in the family tree have special names such as the hypernym relation, the hyponym relation, and the co-hyponym relation. There is a path that connects a general term to the next more specific term. Depending on the direction the name for this path is either called the hypernym relation or the hyponym relation. Mammals are a hypernym of insectivores. In reverse, insectivores is a hyponym of mammals. The hypernym relation is also known by the term superordinate. The hyponym relation is also known by the term is-a relation or subordinate relation. In the case of a hyponym relation between a named entity, such as Groningen, and its mother node city, the leaf node is referred to by the name categorised named entity and the relation is called the instance relation.

The name of the relation between words that are all directly under the same mother node is the co-hyponymy relation. Insectivores and carnivores both belong to the class of mammals and are therefore co-hyponyms. This relation is also referred to by the term coordinate relation. In the metaphor of the family tree they are sisters or siblings.

The differences with the associative relation are plenty. Note that the way we represent taxonomically related words requires directed relations. Fish is a daughter of vertebrates and vertebrates is a mother of fish. In one direction the relation is called hyponymy in the other hypernymy.

We already claimed that the associative relation is looser than the hierarchical relations described in this section. At this point we are able to elaborate on this a little further. While the associative relation is one that only incorporates information about the distance between terms, the taxonomic relations provides information about semantic inclusion: one term subsumes another.

If we take a look at resources that are available for the two types of information: taxonomic relations and associative relations, we see that they are built in different ways. The resources that are available for associative relations are built by conducting association experiments with people. The results are dependent on the group of subjects chosen. One group of people might consider Elvis Presley to be very hip, whereas another group might find Snoop Dogg the coolest thing. Resources that are available for the taxonomic relations are often carefully built by domain specialists. They reflect the decisions taken by a large community. For example, a whale is categorised as a mammal, although it has the looks of a fish. These categorisations are the result of long debates among biologists. Not all domains are well-categorised. Abstract things are a lot less easy to categorise. Often rather ad hoc mother nodes have to be constructed to create a category that brings together a group of lexical items. For example the
mother node *causal agent* brings together *person, agent, nature, supernatural* etc. in WordNet.

### 2.3.3 Synonymy

We have explained that taxonomically related words are words that are close in the hierarchy of meaning such as hyponyms, hypernyms and co-hyponyms. A type of semantically relatedness that we have not discussed so far that is at the very beginning of the scale of similarity is synonymy.

To put it in simple terms there exists a synonymy relation between two words if they share the same meaning. We will give an example of what the lexico-semantic resource WordNet considers to be synonyms. In WordNet (near-) synonymy is represented by means of a so-called synset. Synsets are groupings of synonyms. For example *nature, universe, creation, world, cosmos, and macro-cosm* form one synset. One word can belong to more than one synset, if it has more than one sense. There is another sense of the word *nature*, that is part of the synset that comprises *nature, wild, natural state, and state of nature*.

In literature people have debated about a definition for synonymy. We will give a summary of some views and will explain which notion fits this work best.

Cruse (1986) proposes a scale of synonymy. He argues that since the point of semantic identity, i.e. **absolute synonymy** is well-defined and the other end-point, the notion of zero synonymy, is far more diffuse, a scale of semantic difference is more satisfactory. The definition of absolute synonyms Cruse (1986) gives is the following: “Two lexical units would be absolute synonyms if and only if all their contextual relations (...) were identical.” He then continues with examining an illustrative sample of possible candidates for absolute synonymy. None of the pairs satisfy the criteria. He concludes by stating that “if they exist at all, they are extremely uncommon.” Only in technical domains can one find absolute synonyms, for example *bovine spongiform encephalopathy (BSE), and mad cow disease* are two names for the same thing.

Next on the scale are the so-called **cognitive synonyms**. Cognitive synonyms must be identical in respect of propositional traits, i.e. they must yield the same truth-value, but they may differ in respect of expressive traits. Examples are father-daddy, cat-pussy, infant-baby. Cognitive synonyms arise where certain linguistic items are restricted to certain sentences or discourses. Their cognitive counterparts (synonyms) take their place in other sentences and discourses. Cruse (1986) deals with these restrictions under two headings: (i) **presupposed meaning** and (ii) **evoked meaning**. Presupposed meaning refers to the semantic traits of a lexical item that place restrictions on its normal syntagmatic companions. *Drink* takes for granted an object that has the property
of being liquid. *Grilling* is usually used for raw food such as meat or green peppers, and *toasting* for bread. In the above example the collocational restriction is systematic. In other cases the restrictions can only be described by listing all collocants. These restrictions are referred to with the term idiosyncratic collocational restrictions. An example is the pair *umpire-referee*. Evoked meaning is a consequence of different dialects and different registers in a language. Examples of geographical variety are *autumn* and *fall*, *lift* and *elevator*. Difference in register give rise to cognitive synonyms such as *matrimony* and *marriage*.

From absolute synonyms we went to cognitive synonyms and next we find the *PLESIONYMS (NEAR-SYNONYMS)*. They are distinguished from cognitive synonyms by the fact that they yield sentences with different truth-conditions. Two sentences which differ only in respect of plesionyms are not mutually entailing but there may well be unilateral entailment. *Cruse* (1986) hence categorises hyponyms/hypernyms under the plesionyms.\(^2\)

*Zgusta* (1971) defines absolute synonymy as identity of all three basic components of meaning: *designatum*, *connotation*, and *range of application*. The term designatun refers to a referent of a single word in the extralinguistic world. Synonyms should have agreement in designatun. Connotation refers to the feeling or attitudinal value that a lexical element such as *pass away* distinguishes from *die*. The term range of application refers to the fact that certain words are used in certain contexts. We speak of a *stipend* in connection with a student or researcher, whereas *salary* is used in connection with teachers and other officials. If there is a difference in one or more of the components, words are near-synonyms only.

We have chosen to follow the definition of synonymy given by *Cruse* (1986). When automatically acquiring synonyms from corpora we hope to find cognitive synonyms, we want to find words that are identical in respect of propositional traits, i.e. they must yield the same truth-value, but they may differ in respect of expressive traits. Of course in the event of true synonyms we want to extract those as well, but on the other end of the scale of synonymy we want to limit ourselves to cognitive synonyms and exclude near-synonyms. We have hereby decided for a rather strict notion of synonymy. The fact that we are distinguishing other semantic relations such as the hyponym relation and other related words made us opt for the strict definition of synonymy and not near-synonymy. Also, the fact that we want to apply the synonyms acquired to question answering pushed us in the direction of a rather strict definition. Some of the components such as answer matching and selection require a strict

\(^2\)A problem that arises with substitution tests for synonymy is that they abstract away from potential syntactic or other differences that might affect the substitution test. For example, *ill* and *sick* are synonymous, but because *ill* is only predicative, the substitution is often problematic: *a sick child* vs *an ill child.*
Chapter 2. Lexico-semantic knowledge

Definition. If we were to include near-synonyms we would almost certainly hurt the precision.

However, we do need to extend the definition because of the problem of polysemy. A word that has multiple meanings such as bank naturally gives rise to multiple distinct (cognitive) synonyms. The definition for synonymy adapted to polysemy is as follows: Two words are cognitive synonyms, if there is a sense for both words which allows one word to be substituted for the other in a given sentence without affecting the truth-value of the sentence. Note that we add specifically that there is a sense for which the condition holds. In practice this comes down to the description given by Fellbaum (1998). Fellbaum (1998) notes that WordNet does not entail interchangeability in all contexts. One should speak of synonymy relative to a context. We do not entail interchangeability in all contexts either. Words are synonymous relative to a context.

Figure 2.4 shows what the result is of incorporating synonyms, such as infant-baby, in a hierarchy of related words.

2.4 Available lexico-semantic resources

As explained in Chapter 1, one of the goals of this thesis is to automatically acquire lexico-semantic information to be used in question answering. There are however quite a number of manually constructed resources available. Many of these resources are for the English language, but the famous Princeton WordNet has been extended to European languages (among which Dutch) in the EuroWordNet project (Vossen, 1998).

A question that comes to mind immediately is, if people have struggled for years carefully building lexico-semantic resources, why bother building your own automatically? We need to build further because existing resources are insufficient. We want to apply automatic techniques because it takes much time and effort to build resources manually. A second reason is that language evolves and a manually built resource would have to be updated every once in a while, a time-consuming and expensive enterprise. As a consequence manually built resources normally suffer from low coverage. Moreover, automatic corpus-
2.4. Available lexico-semantic resources

Based methods can be adapted to the domain needed in the current application. The lexico-semantic information can be acquired from the corpus used in the application. It can be updated as often as you like. With every newspaper that arrives from the press in the morning, if that is what you want, or for any domain adaptation. Once in place, the money and time needed is limited.

This is especially helpful to account for lexical variation. For example, Dutch spoken in Flanders is different from Dutch spoken in the Netherlands. Geeraerts et al. (1999) give examples in their study of the lexical variation between Belgian and Netherlandic Dutch for the clothing and football domain. Corpus-based methods for building lexico-semantic resources can be tailored to either two types of Dutch by using texts originating from either two countries. The sem.metrix project\textsuperscript{3} from the University of Leuven aims to measure the structure of lexical variation by using large corpora.

We have, however, used existing resources to evaluate the performance of our system. We realise that there are problems related to evaluating on the resources that you are trying to improve and we will discuss this issue at the end of section 2.5.1. Also, we used these resources as a baseline in our experiments on QA. We will in the next sections (2.4.1 and 2.4.2) give some information about two resources we have used.

2.4.1 EuroWordNet

The aim of the EuroWordNet project (Vossen, 1998) was to build a database of wordnets for English, Spanish, Dutch, and Italian, similar to the Princeton WordNet (Fellbaum, 1998). Princeton WordNet is an electronic resource inspired by current psycholinguistic theories of human lexical memory. Each wordnet in EuroWordNet is structured along the same lines as the Princeton WordNet: synonyms are grouped in synsets, i.e. lists of words that are (near)-synonyms. These synsets are in turn related by basic semantic relations such as the hyponym relation. In addition each meaning is linked with an equivalence relation to a Princeton WordNet synset. Thus a multilingual database is created. We will be concerned with the Dutch part of EuroWordNet only and will refer to it by the term Dutch EWN or simply EWN. Dutch EWN is smaller than Princeton WordNet. According to Vossen et al. (1999), for nouns 56.8\% of the size of WordNet1.5 is reached.

We did a small experiment to see how many of the most frequent nouns in the CLEF corpus were found in EuroWordNet. The CLEF corpus is an 80 million-word corpus of Dutch newspaper text. It is used for the Dutch track of the Cross-Language Evaluation Forum (CLEF), a framework for the testing, 

\textsuperscript{3}http://wwwling.arts.kuleuven.ac.be/qlvl/semmetrix.htm
tuning, and evaluation of information retrieval systems operating on European languages. In the first column of table 2.1 the frequency cut-offs are given. In the second column the number of nouns found in the 80 million-word CLEF corpus are given for each frequency cut-off. In the last column we can see how many of those nouns are found in EWN.

For nouns with a frequency above 1000 92% is found in EuroWordNet. For words down to the frequency cut-off 100 this drops to 68%. It is clear from this table that the coverage of EuroWordNet is not optimal. If we inspect the words above frequency cut-off 1000 that are not found in EuroWordNet, we see that many (78%) of the missing words are proper names, such as Feyenoord, FNV, Fokker, Greenpeace, and Griekenland. Examples of common nouns (with a frequency above 1000 in the CLEF corpus) that are not found in EWN are asielzoeker ‘asylum seeker’, bestuursvoorzitter ‘chairman of the board’, blauwhelm ‘UN peacekeeper’, obligatiemarkt ‘debenture market’, politiemens ‘police person’, but also iemand ‘somebody’, niks ‘nothing’, and ander ‘other’. Some words ended up in the list of nouns due to parse errors, such as vice ‘vice’, which is part of vice-president, and dit ‘this’, dat ‘that’ bet ‘it’ and oud ‘old’. Lastly, there are two multiword expressions: een en ander ‘a couple of things’, and van alles ‘all kind of things’ that are not found in EWN.

### Table 2.1: Number of nouns found in Dutch EuroWordNet at several frequency thresholds

<table>
<thead>
<tr>
<th>threshold</th>
<th># nouns in CLEF</th>
<th># nouns in EWN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1,185</td>
<td>1,095 (92%)</td>
</tr>
<tr>
<td>100</td>
<td>7,741</td>
<td>5,292 (68%)</td>
</tr>
<tr>
<td>50</td>
<td>13,274</td>
<td>7,372 (55%)</td>
</tr>
<tr>
<td>20</td>
<td>27,598</td>
<td>10,217 (37%)</td>
</tr>
</tbody>
</table>

2.4.2 Word association norms

The Leuven Dutch word association norms (De Deyne and Storms, 2008) contain association norms for 1,424 Dutch words. These norms are gathered in a continuous word association task with participants. For each cue, three association responses were obtained per participant. In total, on average 268 responses for each cue were collected. The experiments were conducted between 2003 and 2006 and involved 10292 participating individuals. From this group, 6,329 persons were female, 3,582 were male and 381 persons did not indicate their sexes. The average age was 24 years (SD = 10.55) and was indicated for all but 61 participants. The majority of the participants consisted of 1st-year students at the University of Leuven and at the University of Ghent.
The entire set of stimuli materials consisted of 1,424 words. Some material was taken from previous studies. It contains concepts from various natural categories (fruit, vegetables, insects, fish, birds, reptiles, and mammals), artifact categories (vehicles, musical instruments, and tools), action categories (sports and professions), and a variety of concrete object concepts. The remainder of the items was taken from the semantic categories of weapons, clothing, kitchen utensils, food, drinks and animals. Furthermore, this set was expanded with words corresponding to superordinate concept nouns such as mammal or vehicle. Finally, in the course of the data collection study, new words were added in order to provide norms for the most frequent association responses to the cue words described above.

As the majority of the participants consisted of 1st-year students at the University of Leuven and at the University of Ghent, the data is Flemish. Although Flemish and Dutch as spoken in the Netherlands are highly similar, there are a number of lexical differences. For example, the word *smoutebol* is a typical Flemish word referring to a Flemish type of pastry, fried in oil. The word is unknown to most Dutch speakers. We believe that these cases are exceptional and consider the resource a valid gold standard for Dutch as spoken in the Netherlands. For a study of the lexical variation between Dutch as spoken in the Netherlands and Dutch as spoken in Flanders we refer to Geeraerts et al. (1999).

### 2.5 Evaluating lexico-semantic knowledge

The different types of relations the system proposes all require different evaluation methods. We have introduced associative relations, taxonomically related words, such as hypernyms and co-hyponyms, and lastly synonyms.

Before moving to the discussion of the evaluation of the several types of lexico-semantic relations we would like to say something with regard to the output of the system. For every target word the system outputs a ranked list of words. The system returns several types of lexico-semantic relations: associations, synonyms, and taxonomically related words. We will use the term nearest neighbours to refer to these ranked list of words returned by the system irrespective of the type of semantic relations found.

Let us first discuss the ranked-list output. The ranked list given by the system provides both a rank (depending on the position in the list) and a score attached to each word pair. In Van der Plas and Tiedemann (2006) we have taken the approach advocated by Curran (2003), i.e to evaluate the system’s top-N candidate synonyms, hereby augmenting N gradually. In this way we do not take the scores into account but rely on the ranking only. We will
refer to this method as the RANK-BASED METHOD. We have mostly used the
rank-based method in this thesis. However, in Van der Plas et al. (2008b) and
section 4.5.8 in Chapter 4 we have used the similarity scores attached to the
candidate synonyms to determine a threshold below which candidate synonyms
are no longer taken into account. Words now have varying numbers of candidate
synonyms for every threshold specified. We will refer to this method as the
SCORE-BASED METHOD.

There are several evaluation methods available to assess lexico-semantic
data. Curran (2003) distinguishes two types of evaluation: DIRECT EVALUATION
and INDIRECT EVALUATION. Direct evaluation methods compare the
semantic relations given by the system against human performance or expertise. Indirect approaches do not use human evidence directly, the system is
evaluated by measuring its performance on a specific task. We will refer to
such approaches as TASK-BASED EVALUATION. The direct approaches can be
subdivided in comparisons against gold standards (for example, EWN, synonym
lists, association lists) and comparisons against ad hoc human judgements, i.e.
manual evaluations of the output of the system.

The following sections describe the evaluation framework we have chosen
to evaluate the automatically acquired lexico-semantic information. We will
describe for each type of lexico-semantic knowledge (associations, related words,
and synonyms) how the nearest neighbours are evaluated against gold standards
(section 2.5.1), on the task of question answering (section 2.5.2), and against
human judgements (section 2.5.3).

2.5.1 Gold standard evaluation

Many NLP tasks can be evaluated using a gold standard. In parsing for example one might compare the results of the system with the ones provided in a
manually annotated treebank. For lexico-semantic data several gold standards
are available. We will first give an overview of how gold standards have been
used in literature to evaluate lexico-semantic information. We will then move
to the methods we have chosen to evaluate the three types of lexico-semantic
information: associations, taxonomically related words, and synonyms. We con-
clude by discussing the problems related to evaluating on gold standards in the
last section.

Related work on gold standard evaluation

Rapp (2002) has compared the results of automatic lexico-semantic acquisition
on free word association in addition to the generation of synonyms. They used
the Edinburgh Associative Thesaurus (EAT), a large collection of association
norms by Kiss et al. (1973). For 100 stimulus words they compared the primary 
response from the EAT with the results of their system.

English systems have been evaluated on psycholinguistic evidence such as 
the collected semantic distance judgements on 65 word pairs of Rubenstein and 
Goodenough (Rubenstein and Goodenough, 1965) and modifications of these 
lists (Resnik, 1995; Budanitsky and Hirst, 2001; Weeds, 2003). Also the vo-
cabulary tests of the Test of English as a Foreign Language have been used for 
evaluating similarity systems (Deerwester et al., 1990; Turney, 2001).

Curran and Moens (2002) have compared the nearest neighbours produced 
by similarity measures with thesaurus entries taken from three different thesauri 
(the Macquarie, Bernard (1990); Moby, Ward (1996); Roget, Roget (1911)). 
Weeds (2003) argues that evaluating against these thesauri is problematic be-
cause the neighbour sets extracted should be more akin to WordNet than to 
thesauri such as Roget. Weeds (2003) compares her system to WordNet in a 
WordNet prediction task comparable to work done by Lin (1998a).

For Dutch, we have used Dutch EWN in previous work (Van der Plas and 
Bouma, 2005a; Van der Plas and Tiedemann, 2006). Also, Van der Cruys (2006) 
and Peirsman et al. (2007) have used Dutch EWN to evaluate their systems.

Gold standard evaluation for associative relations

Rapp (2002) has compared the results of automatic lexico-semantic acquisition 
with the primary response from the EAT (the Edinburgh Associative Thesaurus) 
by Kiss et al. (1973) for English. For Dutch we are aware of two resources: the 
Woordassociatie Lexicon (van Loon-Vervoorn and van Bekkum, 1991) and the 
Dutch Word Association Norms (De Deyne and Storms, 2008). We have chosen 
the latter in our evaluations because of its recency and the large size.

Whereas Rapp (2002) looked at 100 stimulus words and their primary re-
 sponses, we have included 1,214 words and all responses given by participants. 
Note that the associations are directed. Broccoli may have green as an associ-
ation, but green might not have broccoli as an association. We have taken this 
into account in our evaluations. We have only used the association directions as 
found in the association norms. We discarded responses with a frequency of 1 
because we have little confidence in these associations. They are highly likely to 
be idiosyncratic. For the top-N associations given by the system for a particular 
word, we calculate how many are found in the Dutch Word Association Norms. 
The result of our evaluation of the candidate associations returned by the system 
will be the average precision of the system with respect to associations found in

\footnote{From the original list of 1,424 words we only considered single nouns. We removed verbs, 
adjectives, and plural nouns. This resulted in a list of 1,214 words.}
the Dutch Word Association Norms.\textsuperscript{5} 

More details about the design of the gold standard evaluation of the associative relations on the Dutch Word Association Norms can be found in Chapter 5, section 5.4.

\textbf{Gold standard evaluation for taxonomically related words}

To measure the semantic relatedness of the nearest neighbours returned by our system we use the EuroWordNet hierarchy (Vossen, 1998). We explained that EWN is organised in the same way as the well-known English WordNet (Fellbaum, 1998). Word senses with the same meaning form \textit{synsets}, and \textit{is-a} or hypernym relations between synsets are defined. Together, the \textit{is-a} relations form a tree-like structure, as illustrated in figure 2.5. The tree shows that \textit{appel} ‘apple’ \textit{is-a} \textit{vrucht} ‘fruit’, which \textit{is-a} \textit{deel} ‘part’, which \textit{is-a} \textit{iets} ‘something’. A \textit{boon} ‘bean’ \textit{is-a} \textit{peulvrucht} ‘seed pod’, which \textit{is-a} \textit{vrucht}.

\textbf{Figure 2.5: Fragment of the is-a hierarchy in Dutch EuroWordNet}

EWN is not a gold standard as such. It does, however, provide us with an approximation of semantic relatedness between words. We will describe how an approximation of semantic relatedness between words can be calculated from EWN.

Every pair of words in the word net is connected by a path. This path can be of varying length. The intuition is that the longer a path is, the less related the terms are. However, it has been proven that pathlength between two terms, more precisely the subtraction of the pathlength from the maximum possible pathlength, is not a good indicator of semantic relatedness between two words.

\textsuperscript{5}With this evaluation method we do not take into account the frequency, nor the order of the associations. We could have used a method that determines the correlation between two ranked lists as in the WordNet prediction task of Weeds (2003) and Lin (1998a), but due to time limitations we have used the method as presented by Rapp (2002) for associations.
2.5. Evaluating lexico-semantic knowledge

This is not surprising since the steps between concepts at the bottom of the taxonomy, where concepts are more specific, represent a smaller semantic distance than at more general top levels of the taxonomy.

There are a number of measures that try to translate the distance in WordNet to a score that correlates well with human judgements. Some try to estimate the distance by accounting for the number of changes in direction in the path (Hirst and St-Onge, 1997) or the location in the taxonomy of the most-specific common subsumer (Wu and Palmer, 1994). Yet another group of measures uses corpus frequencies in addition to the information from the word net to determine the semantic relatedness of words in a word net (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998b). For a comparison of the different techniques see Budanitsky and Hirst (2001).

Of the measures that do not require frequency information Wu and Palmer’s (1994) measure performs best according to Lin (1998b). In our experiments, we have used the measure by Wu and Palmer (1994) precisely because it correlates well with human judgements, and it can be implemented without the need for (sense-tagged) frequency information. Of the measures that do not require frequency information Wu and Palmer’s (1994) measure performs best according to Lin (1998b). In our experiments, we have used the measure by Wu and Palmer (1994) precisely because it correlates well with human judgements, and it can be implemented without the need for (sense-tagged) frequency information. Note that these evaluations apply to Princeton WordNet and judgements for English. Driven by the similar architecture of Dutch EWN and Princeton WordNet we apply the outcome of these evaluations to Dutch and Dutch EWN.

The Wu and Palmer measure for computing the semantic similarity between two words (W1 and W2) in a word net, whose most specific common subsumer (lowest super-ordinate) is W3, is defined as follows:

\[
Sim = \frac{2(D3)}{D1 + D2 + 2(D3)}
\]

Where D1 (D2) is the distance from W1 (W2) to the lowest common ancestor of W1 and W2, W3. D3 is the distance of that ancestor to the root node. The similarity between appel and peer according to the example in 2.5 would be \(4/6 = 0.66\), whereas the similarity between appel and boon would be \(4/7 = 0.57\).

For each pair of a headword and a candidate similar word we calculate the EWN score according to Wu and Palmer (1994)’s measure. If a word is ambiguous according to EWN, i.e. it is a member of several synsets, the highest similarity score is used. Words that are not found in EWN are discarded. The EWN similarity of a set of word pairs is defined as the average of the similarity between the pairs. The system performs well, if the nearest neighbours it finds

---

6 The best measure according to Budanitsky and Hirst (2001) is the measure by Jiang and Conrath (1997). This corpus-based measure uses sense-tagged frequency information. To our knowledge there does not exist sense-tagged frequency information for Dutch words. We therefore applied the measure that correlates well with human judgments and that does not need frequency information.
Chapter 2. Lexico-semantic knowledge

for a given word are assigned a high similarity score according to the Wu and Palmer measure.

We have chosen to give results for the top-$N$ nearest neighbours as in Curran (2003). However, Weeds (2003) and Lin (1998a) have chosen a different strategy, i.e. to calculate the correlation between the ranked lists produced by a WordNet similarity measure and the ranked lists produced by the system.

To summarise, the semantic relatedness between the headword and the top-$N$ nearest neighbours given by the system is computed by measuring the distance in EWN. The result of our evaluation of the nearest neighbours returned by the system will be the average Wu and Palmer score based on EWN.

More details about the design of the gold standard evaluation of taxonomically related words on EWN can be found in Chapter 3, section 3.4.1.

The EWN score, described above, gives an indication of the degree of semantic relatedness in the retrieved neighbours. The fact that it combines several lexical relations is an advantage on the one hand, but on the other hand it is coupled with the disadvantage that it is rather opaque. We would like to decompose this score and see how many of the neighbours found by the system are synonyms, and how many are hypernyms or (co-)hyponyms. We will discuss synonyms in the next paragraph.

To determine the percentage of hypernyms and (co-)hyponyms we again used EWN. For example, to determine if a candidate word is in a hyponym relation with the test word we checked if there is one sense of the candidate word and test word that are in a hyponym relation in EWN. If so, this contributes to the hyponym score for that test word. Note that it is possible for one polysemous word to contribute to the percentages of multiple semantic relations. Therefore, the percentages of the several semantic relations added together can potentially be above 100%.

**Gold standard evaluation for synonyms**

We used the synsets in EWN for the evaluation of the proposed synonyms. In EWN one synset consists of several synonyms which represent a single sense. Polysemous words occur in several synsets. As noted before, the system does not distinguish between the different senses of words. To be able to run a fair evaluation on EWN we have taken the union of all synsets in which a head word occurs as the synonyms for that head word, an approach also taken by Curran and Moens (2002). The result of our evaluation of the candidate synonyms returned by the system will be the average precision of the system with respect to synonyms found in EWN.

Note that this is a very strict evaluation. Curran and Moens (2002), for ex-
ample, have combined near-synonyms from thesauri that are looser than WordNet, such as the Macquarie (Bernard, 1990), Roget’s (Roget, 1911), and Moby (Ward, 1996).

More details about the design of the gold standard evaluation of synonyms using EWN can be found in Chapter 3, section 3.4.2.

Problems related to evaluating on gold standards

Weeds (2003) argues that the system might do badly on the evaluation because of a flaw in the hypothesis which links distribution to semantics. This might be a point when the aim is to evaluate distributional similarity as such. However, since one of our main goals is to be able to extract lexico-semantic information from distributional information, the evaluation on lexico-semantic gold standards is a good starting point.

There is a problem that bothers us more heavily. In the previous section we have explained the motivation for building lexico-semantic resources automatically, basically this is because of shortfalls of the lexico-semantic resources available, of which limited coverage is the most important. Dutch EWN is less than half the size of the English WordNet and hence many words are missing.

We have experienced problems during our evaluations. In Van der Plas and Bouma (2005a) we found that 60% of the words that our system returned as most similar words to a list of 1000 test words from EWN were not found in EWN. We chose to discard words that are not found in EWN and not to count them as errors in our evaluations because they might be valuable additions. Hence these will not affect the scores. However, false negatives, i.e. missing synonymy links between words, when both words are in EWN, but there is no synonymy link between them, do harm our scores. In an evaluation with human judgments (Van der Plas and Tiedemann, 2006) we showed that in 37% of the cases the majority of the subjects judged the synonyms proposed by the system to be correct even though they were not found to be synonyms in EWN. In section 2.5.3 we will discuss these evaluations against human judgements in more detail and mention some of the problems related to this type of evaluation.

A syntactic category for which coverage is minimal are the proper names. As Pasca and Harabagiu (2001) explain regarding Princeton WordNet, ‘the hyponyms of concepts such as composer or poet are illustrations rather than an exhaustive list of instances. For example, only twelve composer names specialize concept composer’. That is to be expected for a manually built resource. The popularity of person names is subject to change. The person that is widely discussed today may not be tomorrow. A manually built resource cannot be updated with regard to the celebrities of the day. This is a serious problem
since the relations between person names are typically very important for a task such as question answering. In the CLEF test set many questions contain person names or ask for answers containing person names. The gold standard EWN does not provide the information we need to assess the quality of the nearest neighbours with respect to proper names.

### 2.5.2 Task-based evaluation

Instead of evaluating the acquired lexico-semantic knowledge directly, one can evaluate how well the acquired lexico-semantic knowledge can be applied in a certain task. In chapter 1 we explained that the work described in this thesis is embedded in the framework of the IMIX project in which our groups is responsible for building a question answering system for Dutch. We have therefore chosen question answering as the task to evaluate the acquired lexico-semantic information on. We will use the testbed for question answering provided by the Cross Language Evaluation Forum for our experiments.

We will first discuss some applications that have been used in related work to evaluate lexico-semantic knowledge. We gave a short summary of the different components of the QA system Joost in section 1.5. We will now explain where we expect the three types of lexico-semantic relations to be most useful. We will start by giving some examples of where associations can be used. Then we will explain where we think taxonomically related words will fit best. The usefulness of synonyms will be discussed in the penultimate section. We conclude by discussing the problems related to task-based evaluation.

#### Related work on task-based evaluation

Examples of task-based evaluations are smoothing for language models (Dagan et al. [1995, 1994]), word sense disambiguation (Dagan et al., 1999; Lee, 1999; Weeds and Weir, 2005) and information retrieval (Grefenstette, 1994b; Ruge, 1992). The PASCAL recognising textual entailment challenge (Dagan et al., 2006) provides an application-independent task that is defined as recognising, given two text fragments, whether the meaning of one text can be inferred from the other. The dataset consists of subsets for seven applications: information retrieval, comparable documents, reading comprehension, question answering, information extraction, machine translation, and paraphrase acquisition. The data covers a broad range of entailment phenomena, many of which are beyond the scope of this thesis.

Automatically acquired lexico-semantic knowledge has also been applied to question answering. Pasca (2004) and Pantel and Ravichandran (2004) present methods for acquiring class labels for instances (categorised named entities),
2.5. Evaluating lexico-semantic knowledge

such as SPSS is a statistical package. Pasca (2004) applies this information to web search, for example, for processing list-type queries. Pantel and Ravichandran (2004) conducted two QA experiments: answering definition questions and performing QA information retrieval (IR). They show that both tasks benefit from the use of automatically acquired class labels.

Task-based evaluation for associative relations

Associative relations group words together according to subject field. It is a rather loose relationship. This is not a type of relation we want to apply to the later stages in the process of answering a question. Stages such as answer matching and selection require rather precise information. We expect that associative relations can be helpful at the stage of passage retrieval. From a small experiment we have done we noted that some questions benefit from associative relations very much. Consider the following question:

(1) Welke bevolkingsgroepen voerden oorlog in Rwanda?
 ‘What populations waged war in Rwanda?’

We expanded the keywords of this question automatically with associations found by the system. Hutu and Tutsi are the second and third associations the system returns for Rwanda. In the first position we find Zaire, which is a less useful expansion, but still the expansions help to find the relevant documents for this question. Expanding a question with associations that in fact constitute the answer obviously helps a lot in finding the right answer.

More details about the design of the task-based evaluation of the associative relations can be found in Chapter 6, section 6.5.

Task-based evaluation for taxonomically related words

A type of semantic relation that we expect to be very helpful for QA at the stage of answer selection and extraction, which is much later in the QA process than IR, is the hyponym relation and specifically the categorised named entities. Since named entities are very important units for QA systems, people often ask information about persons and locations, we expect that the categorised named entities, i.e. NE is-a category, such as Estonia is-a ferry, to be very useful. Consider the example in (2):

(2) Welke veerboot zonk ten zuidoosten van het eiland Utö?
 ‘Which ferry sank southeast of the island Utö?’
Candidate answers that are selected by our system are: Tallinn, Estonia, Raimo Tiilikainen etc. To promote the correct answer Estonia, potential answers which have been assigned the class corresponding to the question stem, i.e. ferry, are ranked higher than potential answers for which this class label cannot be found in the database of hyponym relations. Since Estonia is the only potential answer which is a ferry, according to our database, this answer is selected.

Co-hyponyms are another fruitful source for off-line QA. In off-line QA plausible answers are extracted before the actual question has been asked. An example are the so-called function questions, that ask for a person’s function in some organisation.

Bouma et al. (2005) describe how patterns may be used to extract \langle Person, Role, Organisation \rangle-tuples from the corpus:

\begin{align*}
\text{name}(\text{PER})^{\text{app}} \rightarrow \text{noun} \rightarrow \text{name}(\text{ORG})^{\text{mod}}
\end{align*}

With the previous pattern we extract the tuple \langle Giovanni Agnelli, head, Fiat \rangle from the following text snippet:

chairman Giovanni Agnelli of Fiat

Here, the \text{name}(\text{PER}) constituent provides the Person argument of the relation, the noun provides the role, and the \text{name}(\text{ORG}) constituent provides the name of the Organisation. An important source of noise in applying this pattern to the parsed corpus are cases where the noun does not indicate a role or a function:

colleague Henk ten Cate of Go Ahead

Here, the noun colleague does not represent a role within the organisation Go Ahead.

To remedy this problem, we collected a list of nouns denoting functions or roles from Dutch EWN, and restricted the search pattern to nouns occurring in this list:

\begin{align*}
\text{name}(\text{PER})^{\text{app}} \rightarrow \text{function} \rightarrow \text{name}(\text{ORG})^{\text{mod}}
\end{align*}

While this helps to improve precision, it also hurts recall, as many valid function words present in the corpus are not present in EWN. We expanded the list of function words extracted from EWN semi-automatically with taxonomically related words found in the corpus to get a better recall and yet keep the same precision scores.

More details about the design of the task-based evaluation of the taxonomically related words can be found in Chapter 6, sections 6.4 up to section 6.7.
2.5. Evaluating lexico-semantic knowledge

Task-based evaluation for synonyms

We expect synonyms to be helpful in multiple modules of Joost, for example in query expansion for passage retrieval and for matching between question and answer.

Consider the question:

(3) Hoe oud was Joseph di Mambro toen hij stierf?
    ‘How old was Joseph di Mambro when he passed away?’

and the answer:

(4) Joseph di Mambro was 70 jaar oud toen hij dood ging.
    ‘Joseph di Mambro was 70 years old when he died.’

The answer is a perfect match for the question, but not its surface form. We want to be able to use the information that *dood gaan* ‘to pass away’ is a synonym of *sterven* ‘to die’ to be able to match the question and the answer.

More details about the design of the task-based evaluation of synonyms can be found in Chapter 6, section 6.5 and section 6.6.

Problems related to task-based evaluations

We are aware that there are pitfalls in evaluating components with respect to system performance. The fact that certain components might not benefit from the lexico-semantic information provided does not have to indicate that the information is incorrect or of low quality. It does not even indicate definitely that the information cannot be useful with respect to the task it is applied to. The question answering system under discussion Joost is quite sophisticated. It has lots of heuristics built in that arrive at the same result as the application of lexico-semantic information.

Also evaluating on the questions form the CLEF track are not comparable to using a question answering system with real users. Mur (2006) showed proof that some of the questions in the CLEF track that we use for evaluation look like backformulations. Although Magnini et al. (2004) claim that the questions are made independently from the document collection, the example Mur (2006) gives is rather convincing. The example question she gives is:

(5) Wie was piloot van de missie die de astronomicke satelliet, de Hubble Space Telescope, reparerde?
    ‘Who was pilot of the mission that repaired the astronomical satellite, the Hubble Space Telescope?’.
Chapter 2. Lexico-semantic knowledge

The answer we found was extracted from a sentence in the Algemeen Dagblad of September 19th, 1994, which was formulated as follows:

(6) Bowersox was pilot van de missie die de astronome satelliet, de Hubble Space Telescope, repareerde.
   ‘Bowersox was pilot of the mission that repaired the astronomic satellite, the Hubble Space Telescope’.

The 15 words-long question uses the exact same wording as the sentence holding the answer. If it is indeed the case that questions have been formulated with the document collection at hand, there will probably not be many synonyms nor paraphrases found between question and answer context. In such cases it will be very hard to prove that lexico-semantic information is useful to account for terminological variation in QA.

Related to the problem of backformulations is the fact that the type of questions that are part of the test sets are not as much motivated by what users might want to ask, but rather by what question answering systems are currently able to handle, e.g. factoid questions. The types of questions asked by real users are possibly more complicated and might also contain more lexical variation.

Also, the fact that we are focusing on one application, question answering, has its limitations. It would be interesting to look at inference needs for applications in general. The division of lexico-semantic information as given in the first part of this chapter is mainly motivated by common practice in lexicography. From the perspective of what NLP applications need, we might end up with a completely different taxonomy. We would need to find out what type of information is needed for disambiguation, and what type of information is needed to find out whether a text snippet entails the answer to a question? The PASCAL recognising textual entailment challenge (Dagan et al., 2006) provides an application-independent task. However, the RTE task is for the English language. We are working on the Dutch language. Also, the data covers entailment phenomena that are beyond the scope of this thesis.

2.5.3 Evaluation against ad hoc human judgements

We discussed the shortfalls of the gold standards available. One of the main problems was limited coverage. Ad hoc evaluations against human judgements are not affected by problems of coverage, because people usually have access to a large vocabulary, but one needs to be very careful in setting up the tests.

Another problem with this evaluation technique is the subjectivity of the judgements. It is not an easy task for people to judge semantic similarity let alone associations. Of course looking at agreement between judges can take
2.5. Evaluating lexico-semantic knowledge

away some of the subjectivity. It remains a fact that it is time consuming to run tests with judges, however.

Although human judgments are time consuming, we have ran one ad hoc evaluation to compensate for the shortfalls of the available gold standards discussed above. The evaluation was not done independently. It was used to check the coverage and reliability of the gold standard. By means of a web form presented to subjects we were able to determine the number of false negatives stemming from the gold standard used.

More details about the design and results of the ad-hoc evaluation of synonyms can be found in section 4.4.2.
Chapter 2. Lexico-semantic knowledge