The anatomy of antonymy

Lobanova, Ganna Volodymyrivna

The Anatomy of Antonymy: a Corpus-driven Approach

Anna Lobanova

ISBN: 978-90-367-5875-8 (printed) / 978-90-367-5874-1 (electronic)

Cover image: *The Eye* by Yana Frank
Cover design: Elena Merlo
Printed by Off Page

RIJKSUNIVERSITEIT GRONINGEN

The Anatomy of Antonymy: a Corpus-driven Approach

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 30 november 2012
om 11:00 uur

door

Ganna Volodymyrivna Lobanova

geboren op 1 april 1982
te Odessa, Oekraïne

Promotor:	Prof. dr. L.C. Verbrugge
Copromotor:	Dr. J.K. Spenader
Beoordelingscommissie:	Prof. dr. A. Copestake Prof. dr. A. van den Bosch Prof. dr. P. Blackburn

Посвящается дедушке Мите и бабушке Ане

Contents

CHAPTER 1

Finding opposites automatically: introduction

1.1 Yin and yang, cucumbers and chillies

In the first episode of the TV-show *Chinese Food Made Easy* presenter Ching-He Huang explains the philosophy of *yin* and *yang* in relation to Chinese culture in general and Chinese cooking in particular. "In Chinese philosophy", says Huang, "there are two *opposing* forces, *yin* and *yang*, and the Chinese try to live by a balance of these forces". She illustrates *yin* and *yang* by means of examples like *light - dark*, *male - female*, *cold - hot*. In relation to cooking Huang explains that all ingredients have either *yin* or *yang* qualities and the underlying idea is to balance the two in one's diet. Watermelons, cucumbers and radishes are all instances of *cooling* yin foods while ginger, garlic and chillies are all instances of *heating* yang foods and a proper Chinese dish would contain a perfect balance of the two.

We will refer to words like *cold - hot*, *dark - light*, *large - small*, *man - woman*, *buy - sell* and other pairs that express the opposite of each other as **opposites**. The opposites will be the main focus of this dissertation.

The goal of the work presented in this dissertation is three-fold. First of all, we propose and test three **pattern-based methods** for finding opposites automatically.

We propose methods that automatically generate patterns, that is constructions like [*the difference between <ANT> and <ANT>*] by using sets of seeds, that is well-established pairs of opposites like *rich - poor*, *man - woman*, *buy - sell*. As will be discussed in Chapter 2, automatic extraction of opposites has not been studied as extensively as automatic extraction of other lexical semantic relations. Our goal is to fill in this gap by examining three automatic pattern-based methods for extraction of opposites. Each method relies on different pattern types with varying amounts of syntactic information they contain. These methods have been previously studied in relation to automatic extraction of meronyms (that is, pairs of the type *finger* is part of *hand*) and hyponym-hypernyms (for example, *car* is an instance of *vehicle*). As will be discussed in Section 1.2, automatic identification of opposites is useful for numerous Natural Language Processing (NLP) applications. In this dissertation, we will show that opposites can be found automatically, suggesting which pattern-based methods work best. We will also discuss the impact of our findings on the existing and future work in relation extraction.

Our second goal is to use automatically found opposites for verification of existing classifications of opposites proposed by theoretical linguists. While existing theoretical approaches heavily rely on researchers' intuitions about potential differences between opposites, we take a more objective and methodologically sound approach by using corpus evidence to investigate whether theoretically proposed categories of opposites behave differently in real data usage. We also compare the range of opposites found automatically to those described in theoretical literature, examining overlaps and inconsistencies between them.

Finally, our third goal is to bridge the gap between theoretical linguistic approaches about possible types and functions of opposites, which vary drastically among linguists. We look at the behaviour of opposites found solely automatically in the newspaper corpus. Unlike researchers' intuitions, evidence collected from corpus data is verifiable, it reflects the actual language use and it can be used to study language variations. So far, theoretical linguists have not taken advantage of the opportunities offered by an increased amount of corpus data and computational possibilities, mistakenly assuming that it is not possible to develop a corpus-driven approach to opposites. For example, in a recent journal publication Willners and Paradis suggest that "... there is no method available for identifying types of relation [*between opposites*] correctly. For instance, it is not possible to tell the difference between antonyms [*opposites*], synonyms [*rich - poor*] and other semantically similar word pairs [*weak - short*]." (Willners and Paradis [2010], pp.18). The goal of this dissertation is to show that it *is* possible to auto-

matically identify opposites, differentiating them from synonymous and semantically similar words. Moreover, our aim is to demonstrate that it is of essential importance to use computational means for developing theoretical classifications of opposites, their types and functions.

1.2 Why is automatic extraction of opposites useful?

Being able to reliably find opposites is useful for many areas of linguistics. In computational linguistics, an increasing number of NLP applications relies on opposites. So far, such applications used mostly hand-crafted computational lexical resources like the well-known and widely used Princeton WORDNET for English (Fellbaum [1998]). It contains not only opposites, but also synonyms (pairs like *rich - wealthy*), meronyms (pairs like *finger - hand*), hyponym-hypernyms (pairs like *car - vehicle*) and other relationships. However, manually constructed resources have limited coverage and do not include many good pairs of opposites. Moreover, such resources are expensive (time- and cost-wise) to build, to maintain and to improve as more and more pairs have to be added. For example, extra steps need to be taken to include domain-specific opposites like *dextrorotation - levorotation* and newly lexicalized opposites like *paper - digital*. A further problem is that such resources are mostly created for English. As a results, algorithms that rely on hand-crafted resources often cannot be extended to languages other than English. Finding opposites automatically would resolve these shortcomings since an automatic method can be applied to different languages, allowing to use the results to augment and verify existing lexical resources for English but also other languages, for example, Dutch.

Automatically found opposites can also be used to automatically identify the discourse relationship of Contrast (Marcu and Echihabi [2002], Spenader and Stulp [2007]). For example, knowing that *tall* and *short* are opposites will help to automatically identify contrast in the sentence "*John is tall but Bill is short.*". Opposites are particularly useful for identification of contrast relationships when such discourse markers of contrast as the connectives *but*, *although* and *while* are missing. For example, knowing that *open* and *closed* are opposites will help to identify contrast in "*Everyone assured us the offices would be open on Saturday. They were closed.*" (Spenader and Stulp [2007], p.2).

Opposites have also been used for automatic identification of contradictions and paraphrases (de Marneffe et al. [2008], Voorhees [2008]). Namely, de Marneffe et al.

[2008] argue that opposites provide a very reliable cue for identification of contradictions, as in "*going to war to establish peace*". Another example of contradictions that can be identified by means of opposites are sentences like "*Capital punishment is a catalyst for more crime*" and "*Capital punishment is a deterrent to crime*" (examples from de Marneffe et al. [2008], p.1041). Knowing opposites is also useful for automatic generation of paraphrases such as "*Mars may not be lifeless*" and "*Mars might have life*".

Also humour often contains oppositions and contradictions indicated by opposites. For example, the satire effect in "*Always try to be modest and be proud of it!*" is due to the contrast between opposites *being modest* and *being proud*. Identification of such opposites can help to increase accuracy of automatic humour recognition systems (Mihalcea and Strapparava [2005]). As Mihalcea and Strapparava [2005] point out, lexical resources like WORDNET are far from complete in their coverage of opposites for this task. Automatic acquisition of opposites, on the other hand, offers a more direct approach as automatic pattern-based methods described in this dissertation can be applied to find opposites in a domain-specific corpus, for example, the same corpus of humorous texts used in Mihalcea and Strapparava [2005].

Also Mohammad et al. [2008] suggest that opposites play a crucial role in multi-document summarization, especially when texts contain opinions. Opposites have also been used for creating an emotion lexicon (Mohammad and Turney [2010]). Finally, a list of automatically found opposites can also be applied as a filter to improve the performance of automatic techniques for synonym, hyponym and meronym extraction (Lobanova et al. [2009]), where opposites form a notorious problem (Lin et al. [2003]).

In summary, more and more NLP applications can profit from knowing which pairs are opposites, thus, it is necessary to develop methods for finding opposites automatically.

1.3 Research questions

In relation to work that has been done by computational linguists in automatic relation extraction, we address the following research questions:

- Can opposites be found automatically?
- Which pattern types perform best? Does syntactic information improve precision and recall?

- Is automatic extraction of opposites equally successful, that is productive, for finding opposites expressed by different syntactic categories, namely, adjectives, nouns and verbs?
- How does the size of the corpus affect the results of a pattern-based method for finding opposites? [1]
- How does the genre of the selected corpus affect the results?

In relation to work that has been done by theoretical linguists on classification of the types and functions of opposites, we address the following research questions:

- Do automatic methods find the same types of opposites that have been extensively studied in theoretical linguistics?
- Are automatically generated patterns for finding opposites qualitatively different from manually constructed patterns that were used to study opposites in corpus-based studies? Do they find different types of opposites?
- Are antonyms found automatically the same as the ones described in theoretical literature? That is, how wide is the range of antonyms (in relationship to their types) extracted automatically?

1.4 Dissertation overview

The rest of this dissertation is organized as follows. **Chapter 2** can be divided into two themes. First, it introduces the concept of antonymy, providing a thorough description of the existing theoretical approaches to antonymy, giving examples of various classes of opposites and discussing their limitations and implications for the current work. Chapter 2 also presents existing work on pattern-based methods for automatic acquisition of such lexical semantic relations as hyponymy and meronymy, and it introduces computational work that has been done in relation to antonymy. A combination

[1] While the first corpora introduced in the 90's were very small by today's standards (think of the Brown corpus that consisted of one million words Kučera and Francis [1967]), today more data is available for computational linguists ranging from one hundred million words to gigabytes of data available thanks to such resources as Wikipedia and Google. However, the trade-off is that processing more data requires more computational power. In Chapter 4 we address the question of whether larger corpora lead to better results, that is higher precision and higher recall, or whether once a sufficient amount of data is used, the results do not differ substantially and using larger data sets is redundant.

of the previous accounts on automatic relation extraction and corpus-based work done on antonymy has laid the foundation for the pattern-based methods we propose and test in this dissertation.

Chapter 3 explains the methodology used throughout our work for the evaluation of automatically found candidate pairs. We discuss existing difficulties in assessing the quality of the results and propose a framework for classification of found pairs based on manual evaluation, as well as the usage of existing computational lexical resources and dictionaries.

Chapter 4 presents the first pattern-based method for finding opposites automatically by means of *automatically generated* ***textual patterns***. Textual patterns, or surface constructions like [*the difference between <ANT> and <ANT>*] capture the linear ordering of words in a sentence and do not contain any syntactic information about them. This means that although we use seed sets expressed by different syntactic categories, the instances of *rich - poor* (noun - noun) and *rich - poor* (adjective - adjective) pairs will not be disambiguated by the algorithm. This chapter also investigates the role of *the genre of the corpus* by conducting the same experiments on a corpus of newspaper texts and a corpus of encyclopaedia texts.

Chapter 5 presents results from the second pattern-based method that uses *automatically generated* ***part-of-speech patterns***, for example, [*the difference between <ANT>/ ADJ and <ANT>/ADJ*]. Such patterns preserve the linear ordering of words in a sentence but they contain information about their syntactic categories and therefore they disambiguate instances of *rich - poor* (noun - noun) from instances of *rich - poor* (adjective - adjective).

Chapter 6 presents the third pattern-based method for finding opposites that uses *automatically generated* ***dependency patterns***, that is patterns that abstract away from the surface structure and capture syntactic relations between words. Dependency patterns can deal with long dependencies, that is, sentences in which opposites co-occur too far away from each other so that they cannot be found by means of patterns that preserve linear ordering of words. They can also disambiguate between opposites expressed by different part-of-speech categories extracting opposites of a target syntactic category defined by the seed set. In this chapter we also discuss implications of our findings for an ongoing debate in the literature on automatic relation extraction as to whether dependency patterns (Chapter 6) outperform surface patterns (Chapter 4 and Chapter 5).

Finally, **Chapter 7** draws general conclusions in relation to the results on auto-

matic extraction of opposites by means of patterns presented in earlier chapters. We also discuss implications of the results for the existing theoretical and corpus-based approaches to antonymy, suggesting directions for the future work.

CHAPTER 2

Introducing opposites

This dissertation is about automatic extraction of *opposites*, that is, pairs like *good - bad*, *boy - girl*, *buy - sell*, and so on[1]. Without realizing it, we continuously encounter such words in every day life. When entering a shop, a sign will indicate whether *to push* or *to pull* the door. Traffic lights indicate whether *to stop* or *to go*. The elevators take us *up* and *down*. While antonymy is the "most readily apprehended" relation between senses of words by native speakers including children (Cruse [1986], p.197), the linguistic understanding of this relation and its functions is far from complete. This makes automatic extraction of opposites a much more difficult task than identification of such relations as hyponymy and synonymy.

In this chapter, we present and analyse existing approaches to classification of antonymy and its types. First, we discuss theoretical approaches to antonymy, describing a vast number of categories proposed by different linguists, and their main shortcomings. The latter include inability to provide a clear definition of what antonymy is and to give linguistic tools for distinguishing opposites from non-opposites. We dis-

[1] Parts of the material in this chapter have been published as Anna Lobanova, Tom van der Kleij and Jennifer Spenader [2010] Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: *International Journal of Lexicography*. Vol 23, pp.19-53.

cuss how corpus-driven work on antonymy can play a role in classification of opposites and outline how our results can contribute understanding to the existing theories on antonymy. In particular, we argue that automatically found pairs provide new evidence based on real data usage about the similarities and differences between well-established opposites like *rich - poor* and non-typical pairs like *city - farm*.

Next, we introduce corpus-based studies on antonymy and its functions in discourse. These studies lay the foundation for the current work. But while their main focus is on examining typical well-established opposites, our goal is to find a wide range of all types of opposites, many of which are not covered by theoretical approaches. Results obtained from real data can be used to study behavioural patterns of different types of already-known opposites. Real data is also an excellent source for finding novel pairs that have not been previously investigated. The final part presents previous work on relation extraction in computational linguistics research as well as previous computational work done in relation to antonymy. This part will give a thorough understanding of why we chose specific pattern-based methods and how and why they have been used in the past for automatic extraction of relations other than antonymy.

2.1 Confusion about antonymy

As will become clear in the next few pages, antonymy as a relation causes a great amount of confusion and disagreement among scholars. This makes the study of automatic acquisition of opposites difficult to conduct as the range of possible opposites we can find is not well-defined. Before studying opposites and their functions, one has to decide whether to use the term antonymy in its broad or narrow sense. In its broadest sense, antonymy covers a wide range of word pairs expressed by different part-of-speech categories as long as these words express the opposite of each other. This approach is taken in this dissertation. In its narrow sense, however, antonymy is a relation that holds between a small number of adjective - adjective pairs only. Section 2.2 presents an overview of the existing approaches to antonymy and explains why we view antonymy in its broadest sense.

Usually, the term *antonymy* is used to refer to *binary* opposition only. In particular, it is said that antonymy holds between two words that denote the opposite poles along a certain scale. For example, the opposites *hot* and *cold* refer to the opposite poles on the scale of TEMPERATURE, and the opposites *tall* and *short* refer to the opposite poles on the scale of HEIGHT. However, antonymy covers a much wider range of pairs, and

such pairs are not necessarily binary. Think of the pair *to listen - to speak*. While these words belong to a *multiple-member* category that also includes members *to read* and *to write*, one can easily think of an example sentence in which this pair is antonymous. For example, in the sentence '*With this walkie-talkie you can either listen or speak*' the words *to listen* and *to speak* refer to two opposite and *mutually exclusive* actions.

Although everyone is able to recognize opposites, often on the intuitive level, especially when it comes to typical opposites, also referred to as *canonical* (for example, the pair *rich - poor*), none of the existing definitions of antonymy provides adequate characteristics that can be used to separate opposites from non-opposites. As a result, even the widely-used classifications of antonymy and opposites, such as the classification proposed by Cruse [1986], are unable to deal with non-typical pairs (for example, *to read - to write*, *city - farm*), as they do not provide neither a clear definition of what antonymy is and *what it is not* nor reliable linguistic tools for distinguishing opposites from non-opposites. Thus, such approaches cannot be applied to evaluate these pairs.

2.1.1 'Antonyms' or 'opposites'?

Before going any further, it is important to clarify what is meant by *opposites* and *antonyms* throughout this dissertation. Previous accounts on antonymy have made multiple distinctions among types of opposites. Some prefer to keep the term *antonyms* to refer to a specific sub-class of opposites expressed by gradable adjectives only (Lyons [1968], Lyons [1977], Cruse [1986]), others argue that the term *antonyms* should be used more generally and cover various types of otherwise-known *opposites* expressed by adjectives, nouns, and verbs (Jones [2002]). Unfortunately, choosing one term over the other does not help to solve any of the existing problems in giving a clear definition of the relation itself nor does it add any further understanding as to how to distinguish opposites from non-opposites in difficult cases.

Murphy [2003] proposes to use the terms *antonyms* and *opposites* interchangeably, suggesting that all antonymous pairs share core antonym properties and will be recognized as such by any native speaker. In our studies, we are interested in finding all kinds of opposites, so that our results can be used in various Natural Language Processing applications. For that reason, we use the term *opposites* in its widest sense, referring to any kind of binary and non-binary pairs that indicate opposition in meaning. Illustrating opposites is an easier task than defining them mostly because the notion of oppositeness itself is difficult to delimit.

We will now discuss in detail categories of opposites proposed by theoretical linguists. We show that some of the suggested categories overlap, and others have a lot of exceptions that do not follow the generalizations. We argue that the problems with theoretical classifications, which are mostly based on the researchers' intuition, can be overcome if we take a more data-driven approach to study opposites.

2.2 Theoretical view on antonymy

Theoretical research has focused on semantic or logically based classifications of opposites. The most well-established classes of opposites are the ones expressed by adjectives. Starting with Lyons [1977], a fundamental distinction is made between ***gradable*** and ***non-gradable*** opposites.

2.2.1 Well-established opposites: gradable adjectives

Gradable opposites, also known as *contraries*, include pairs *wide - narrow*, *cold - hot*, *small - large*. All such pairs are expressed by adjectives, and describe the opposite directions along a given scale representing degrees of a certain relevant property. As such, they can be modified by modifiers such as *very*, *slightly*, *fairly* and so on (Cruse [1986]). For example, the pair *wide - narrow* denotes the opposite directions on the scale BREADTH, and can be described as *very / fairly / slightly wide* or *very / fairly / slightly narrow*. It occurs in comparative and superlative constructions, for example, "*The Tower Bridge is wider than London Bridge*".

Lehrer and Lehrer [1982] distinguish a subclass of *perfect opposites*, that is gradable adjectives that are placed on the scale symmetrically, for example, *cold* and *hot* as opposed to *cold* and *tepid*. The scale TEMPERATURE, evoked by opposites *hot* and *cold* contains a middle point, the pivotal region, that cannot be referred to by either of the words, or any other lexical item (Cruse [1986]). As a result, it is possible to be *neither cold nor hot*, *neither wide nor narrow*, *neither long nor short*. The scale itself can be thought of as having a zero point, which corresponds to the absence of the evoked property, extending indefinitely to the direction of "more of" the property. However, the scale on which gradable opposites operate is always relative to the entities the opposites refer to. For example, although *small* and *large* operate on the absolute scale SIZE, a small bear is larger than a small cat. Similarly, a narrow valley is wider than a

narrow street. In other words, *small* and *large*, *narrow* and *wide* are always interpreted as *being smaller/larger* or *wider/narrower* in relation to the referent.

As a rule, pairs of gradable opposites contain a marked and an unmarked term. The unmarked terms are neutral in questions, so that when unmarked terms are used in questions like "*How tall is the player?*" or "*How large is the TV-set?*", the speaker does not make any suggestions as to the height/size of the referent. On the contrary, when the marked terms are used, for example, "*How short is the player?*" or "*How small is the TV-set?*", the speaker suggests that the player is not tall or that the TV-set is small. The unmarked terms are also used in nominalizations, for example, *warmth*, *height*, *width*. Only the unmarked terms can be used in comparative constructions such as *twice as old/wide/tall/large*. Different approaches describe different criteria for identifying marked and unmarked terms among gradable opposites, nevertheless, identifying a marked term in a pair is not always an easy task. Because of that, it is not possible to use markedness as a tool for identifying good opposites.

This led Cruse [1986] to subdivide gradable opposites into further subtypes. However, such fine-grained distinctions have been criticised for being too subjective, reflecting a researcher's intuition rather than real-world examples found in corpora. For example, Cruse makes a distinction between opposites *rude - polite* and *happy - sad* suggesting that "*John is rude but he is more polite than Bill*" is an acceptable sentence whereas "*John is sad but he is happier than yesterday*" is not. Jones, however, argues that according to his intuition both of these sentences are acceptable in English (Jones [2002], p.16). This demonstrates that researchers' intuitions are biased and it is necessary to use a more reliable approach, for example, the one taken in this dissertation where we use real usage data to find valid pairs of opposites.

As has been already said, gradable adjectives provide the most typical examples of opposites, as they exhibit mostly symmetric semantic contrasts. Non-gradable opposites pose a much bigger challenge for linguists, especially pairs expressed by syntactic categories other than adjectives. The further confusion arises from the fact that while Lyons [1977] and Cruse [1986] discard non-gradable opposites from being antonymous, Kempson [1977] suggests that only non-gradable opposites are truly antonymous. Such discrepancies in the views on antonymy underline how difficult it is to define this relation. At the same time, the ongoing debate makes antonymy a fascinating topic and an appealing subject for corpus-driven research which relies on the real data rather than on scientists' intuitions.

2.2.2 *Difficult cases: non-gradable opposites*

Non-gradable opposites provide an interesting mix of binary and non-binary pairs expressed not only by adjectives but also by nouns and verbs. The lack of scales makes non-gradable opposites difficult to classify. Because of that, there is a lot of confusion and inconsistencies between different approaches to their classification. Since we are interested in finding opposites expressed by different part-of-speech categories, we discuss non-gradable opposites in detail.

2.2.2.1 *Beyond adjective - adjective opposites: binary opposites*

The easiest category of non-gradable opposites to grasp is the category of *complementaries*, which is recognized by many different authors. Traditionally this term is used to refer to binary pairs that exhaustively bisect a domain so there is no middle point, for example, *dead - alive*, *man - woman*, *to fail - to succeed*. As a result, unlike gradable opposites, complementaries are *mutually exclusive* and cannot be used with degree modifiers. For example, the pair *married - single* is mutually exclusive because *X is married* entails *X is not single* and, vice versa, *X is single* entails *X is not married*. Cruse [1986] further argues that it is also not possible to be *neither married nor single* nor is it possible to be *very/extremely single* or *very/extremely married*. However, the latter argument exposes one of the main weaknesses of the approaches that do not take real data into account, as it is easy to find counterexamples in which *very married* and *very single* are perfectly acceptable. For example, a post on one of the dating sites[1] contained the following *''2 of the men I met through this site, were "very" married though they declared to be single. Absolutely unacceptable!! Has this ever happened to you?*". The reply said "*I am married, but "very" single..... is that acceptable?*". Although the modifier *very* is taken into quotes in both cases to show that the intended meaning is not literal, any native speaker will judge these utterances as perfectly acceptable sentences in English.

Another weakness of the definition used above is the notion of the domain. Cruse [1986] suggests that the denial of one of the terms entails the assertion of the other, which in itself evokes the domain. For example, asserting that *X is not a female* entails that *X is a male*, evoking the domain of 'human beings'. But, as Cruse notes himself, using entailments as a tool for domain identification does not always work, as it is easy to find exceptions in most of the cases. For example, zombies are neither dead nor

[1] http://www.connectingsingles.com

alive. In a similar vein, snails possess both male and female reproductive organs. As an explanation, Cruse suggests that there is a continuum between contraries, that is gradable opposites expressed by adjectives, and complementaries, that is non-gradable opposites expressed mainly by adjectives, verbs, and entailments can only be used to identify clear-cut pairs of opposites, for example, contraries *tall - short*, and complementaries *to fail - to succeed*. Intermediate pairs, on the other hand, are difficult to categorize. However, because Cruse's approach relies on so many exceptions, it is not appealing to accept such classification.

Further, the fact that it is possible to encounter examples like "*I was more dead than alive.*" or "*I met Katy when she was very pregnant.*" led some linguists to question the whole concept of opposites' non-gradability and, consequently, the necessity to make a distinction between opposites based on their gradability at all. Cruse again proposes to treat opposites in such examples as a special case of *gradable complementaries*. According to him, pairs like *dead - alive* have at least two senses so that in one sense they are mutually exclusive complementaries and in the other sense they are gradable contraries. Jones [2002] doubts the necessity to make theoretical distinctions based on gradability, suggesting that examples from corpora show that the two categories behave similarly and instead of postulating each such case as an exception or a further sub-class, it is better to remove the distinction between opposites that is based on their gradability.

Interestingly, there are also non-gradable opposites that exhibit some properties of gradable opposites. For example, the verbs *to love - to hate*, *to approve - to disapprove*, *to please - to displease* behave similarly to gradable adjective opposites, e.g. *happy - sad*, *cold - hot*. They represent the opposite directions on a scale ADMIRATION; they are not mutually exclusive: one can neither like nor dislike something/someone (although it is not possible to measure the degrees of liking or disliking); and they can be modified as in "I quite like it!", "I absolutely love it!". Already here we would like to point out that fine-grained approaches to antonymy in its narrow sense lead to many exceptions, most of which are taken from the real-world examples. We take this as evidence to refer to our findings as antonymy in its most broad sense. We expect to find many real-world examples of opposites that fall outside of the classes of opposites based on the intuition of the researchers.

2.2.2.2 *Beyond adjective - adjective opposites: non-binary pairs*

So far we have discussed mostly opposites expressed by adjectives and to a lesser extent by verbs and nouns. While adjective opposites are usually binary pairs, many verb opposites are triplets. For convenience, we will refer to pairs that contain more than two members as *non-binary opposites*. For example, *be born - live - die*, *invite - accept - turn down*, *try - succeed - fail*, *attack - defend - submit* are all examples of non-binary opposites. Because there are differences among such triplets and no unified properties can be established, Cruse [1986] proposes to subcategorise them into four further classes. The problem is that even with four subcategories, it is difficult to draw generalizations that would cover all examples of the given subtype. For example, the triplets *be born - live - die* and *learn - remember - forget* both belong to the subtype *reversives*. The outer pairs, for example, *to die* and *to be born* represent change in opposite directions, in particular, leaving / entering the life, and the combination of the three verbs indicates opposition between continuance of one state and the change to the other. It is difficult to analyse the second triplet in the same vein as it is possible *to learn* and *to forget* in turns, or even to learn something and to forget something else at the same time, without representing change in opposite directions. Given that reversives are a subclass of *directional* opposites, that is, words that denote movement in opposite directions, and that there is a further subdivision of reversives into *independent* and *restitutive* does not help to categorize all pairs. This suggests that approaches based on the intuition lead to an overproduction of necessary categories of antonym types. Again this implies that a different approach that is not based on the distinction between gradable and non-gradable opposites can yield a better understanding of antonymy.

What is completely missing in theoretical classifications of antonymy is how such pairs, including typical opposites as well as exceptions, behave in corpora, and whether they differ from one another with respect to their usage in discourse. Before moving on to address these questions, we discuss theoretical classifications of opposites expressed by nouns. Unlike adjectives and verbs, many opposites expressed by nouns are non-binary and have more than three counterparts. Consequently, these pairs present the most disputed categories, making it important to understand the viewpoint on such cases.

Lyons [1977] identifies *converses*, that is, pairs like *parent - child*, *father - son*, as well as *to buy - to sell*, *above - below*), in which it holds that if *X is q to Y* then *Y is p to X* and, vice versa, *if Y is p to X* then *X is q to Y*. Cruse [1986] designates such pairs as relational opposites called *converses*. Converses consist of pairs that denote direction

of one entity in relation to another so that "X *is a descendant of Y*" (for example, "*X is a child of Y*") is the opposite of "*Y is the ancestor of X*" (for example, "*Y is the father of X*"). The opposition between such relational pairs is directional, in particular, the *father* is *above* and the *son* is *below* and the property is passed down from the father to the son. Other examples of noun converses include *master - servant*, *guest - host*, *teacher - pupil*, *predator - prey*. Less obvious examples are the pairs *husband - wife* and *aunt - uncle* where it is difficult to think of a directional opposition, even though Cruse himself argues that "... it is not difficult to think of husband and wife as facing one another, as it were, along the marital axis." [1986, p.232].

Direction in opposition is one of the key properties of opposites identified by Cruse. He points out the necessity to identify direction for every pair. For example, *antipodals*, that is opposites that express the two extremes in the opposite directions, include pairs like *top - bottom*, *maximum - minimum*, *attic - cellar*. They are instances of antipodals as they identify the extremes of directions *upwards* and *downwards*. Also the pair *source - mouth* is an antipodal as it encapsulates the notion of upward and downward streams.

The other key property for a pair to be recognized as opposites according to Cruse is its inherent binarity, or "... an ineluctable 'two-ness' in the relationship" [1986, p.258]. Inherent binarity reflects binarity of unidimensional scales, whose axis can never have more than two extremes. Inherently binary opposites are located along the scale symmetrically away from the middle point or the pivotal area. Also if negation of one term asserts the other, this indicates inherent binary opposition between them.

2.2.2.3 Multiple incompatibles

While Cruse discards non-binary pairs as opposites, others argue that non-binary opposites exist and that they form a separate category called *multiple incompatibles*. This category includes, for example, the closed set of the *seasons of the year*, in which *winter* is incompatible with *summer*, *fall* and *spring*. Lyons (1977) argues that *military ranks* are a case of *ranked multiple incompatibles* where *a general* is incompatible with *a private*. He also classifies sets like *man - woman - girl - boy* as instances of *orthogonal opposition*, another type of opposition where each member of the set is in opposition with two other members. For this example, *man* is opposed to *boy* and *woman*, and *girl* is opposed to *boy* and *woman*. Others either ignore such cases (Palmer [1976], Jackson [1988]) or treat such pairs as co-hyponyms, that is sister nodes that

share the same hypernym, and do not acknowledge this potential oppositional meaning. However, for most of the NLP applications that rely on opposites (see Chapter 1 for details), knowing that *summer* is the opposite of *winter* is as relevant as knowing that *rich* is the opposite of *poor*. This implies that for us it is as important to find binary opposites as it is to find multiple incompatibles. In turn, if our automatic methods do not find multiple incompatibles, this will provide evidence for approaches that do not recognize them as opposites. In this case, it will be possible to suggest that mutual incompatibles are co-hyponyms rather than opposites. None of the previous work on antonymy has addressed this issue from a computational perspective. Our findings will be therefore of primary importance for shedding light in this ongoing debate.

None of the approaches described so far have taken the *context* of opposites into account. Murphy [2003] presents a contextual approach to classification of opposites. She argues that the preference for candidate opposite pairs could change depending on the context. For example, in a neutral context *sweet - sour* seem to be better opposites in English than the pair *sweet - bitter*. Depending on the context, however, *sweet* can have other opposites, including *salty* (in relation to popcorn), *dry* (in relation to white wine), *bitter* (in relation to linctus (liquid medicine)) and so on. Further, the preference for a certain opposite candidate can vary across languages. For example, in a neutral context, the opposite of *sweet* in Japanese is *karai* ("spicy-hot and/or salty") and in Korean it is *bitter* (Backhouse [1994], cited in Murphy [2003], p.173). While previous accounts could not explain variation preferences across languages, a context-dependent approach can explain such differences as it assumes that the range of opposites for any particular word exceeds its counterparts in a neutral context and given specific contexts, contextual cues may override semantic cues and preferences found without the context. Our automatically generated patterns are acquired from the sentences in which opposites co-occur, providing a minimal context for candidate pairs.

2.3 Corpus-based approaches to antonymy

2.3.1 Sentential co-occurrence of opposites

While theoretical linguists study taxonomies of opposites based on their semantic properties, psycholinguistic studies of opposites suggest that these distinctions do not play a role in the way opposites are represented in the mental lexicon. Stimulus-responses for word association tests have shown that a subset of adjectival opposites have fea-

tures unique among lexical relations: priming participants with one member of the pair leads them to respond with the other member of the pair (Deese [1964]; Deese [1965]), suggesting that adjectives are learnt together as opposites. This is in part because antonymy is so ubiquitous with adjectives, and it has even been argued to be the organizing semantic relation for this class of words in the mental lexicon (Deese [1964]; Deese [1965]). The set of opposite pairs that display this type of response includes pairs found in the traditional categories of gradable and non-gradable binary opposites. Such a response is used as further evidence of their canonicity, or typicality. Multiple incompatibles were not part of the study. Deese concluded that such antonymous adjectives are strongly associated because they share identical contexts and as a consequence they can be substituted for one another.

This idea is now known as the *Substitutability Hypothesis*. Charles and Miller [1989] tested this hypothesis by extracting sentences that contained one of the adjectives from the pairs *weak - strong* and *public - private* from the one million word Brown Corpus of English (Kučera and Francis [1967]). They then created experimental materials by removing the adjectives from the sentences and leaving a blank. They asked participants to fill in the missing adjectives in either the full sentence or part of it. If opposites from the same pair were mutually interchangeable, participants would have no preferences as to the choice of adjectives, and would fill in each equally as often. However, the results showed that in many contexts, only one of the adjectives was appropriate. For example, *coffee* can be described by the adjective *strong* but not as readily by its opposite *weak*. Similarly, a *hospital nurse* is likely to be modified by the adjective *public* rather than *private*. This was taken as evidence against the Substitutability Hypothesis.

Instead, Charles and Miller [1989] argued that canonical adjectival pairs are learned as such because they co-occur in sentences more often than would be expected by chance, an idea they called the *Co-occurrence Hypothesis*. The idea that opposites co-occur with each other within a sentence significantly more often than is predicted by chance has become the fundamental assumption in all corpus-based work on antonymy, including studies presented in this dissertation.

Originally, Justeson and Katz [1991] tested the Co-occurrence Hypothesis on a small corpus, examining the frequencies of intrasentential occurrences of adjectival opposites in the Brown Corpus (Kučera and Francis [1967]). They confirmed that a set of adjectival opposites co-occurred together significantly more often than sets of random adjectives. Moreover, in many sentences adjective - adjective opposites co-

occurred in specific textual patterns like [*between* <*ANT*> *and* <*ANT*>] and in these patterns opposites could be substituted for one another. This result led to the conclusion that while frequent co-occurrence may be a characteristic of canonical opposites, it is not sufficient to establish the relationship because many lexically related words co-occur together significantly more often than chance. Instead, co-occurring in certain intrasentential patterns like *adjective - conjunction - adjective*, or in parallel constructions where an opposite pair modifies two identical nouns, is necessary for establishing the strong lexical association found in psycholinguistic tests.

Because Charles and Miller looked at sentences with only one of the two opposites, the contexts were not always interchangeable. Justeson and Katz demonstrated that in those sentences, in which both opposites co-occurred together, they could be substituted for one another. These two hypotheses, the Substitutability Hypothesis and the Co-occurrence Hypothesis, have been the foundation of the corpus-based work on antonymy.

2.3.2 Co-occurrence of opposites expressed by nouns and verbs

Note that initially the Substitutability and Co-occurrence hypotheses were used only to study antonymy expressed by adjective - adjective pairs. Fellbaum [1995] conducted the first large-scale corpus work that looked at a wider class of opposites that included also nouns and verbs. She looked at the co-occurrences of nominal and verbal opposites in the Brown Corpus and found that opposites in both groups co-occurred in the same sentence significantly more often than chance. However, unlike adjectival opposites, they did not co-occur in parallel constructions or specific textual patterns with the same regularity as adjective pairs. In fact, intrasententially co-occurring antonymous nouns often differed in their number (singular/plural) while co-occurring antonymous verbs frequently had different subjects and were in different tenses. In relation to our work, this implies that automatic pattern-based methods can perform better at finding opposites expressed by adjectives rather than nouns or verbs. However, as Fellbaum notes herself, when noun - noun and verb - verb opposites were found in patterns, the patterns were the same as the patterns filled by adjective-adjective opposites. This suggests that given a larger enough corpus, patterns can be automatically identified to the same extent for all three part-of-speech categories.

2.3.3 *Cross-categorical opposites*

Fellbaum [1995] also looked at the intrasentential co-occurrences of morphologically related word pairs that express semantic opposition but do not belong to the same syntactic category, for example, pairs such as *to begin* (V) and *endless* (Adj), or *death* (N) and *to live* (V). Again, these cross-categorical antonym pairs co-occurred significantly more often than chance in the same sentence, suggesting that opposites do not have to belong to the same syntactic category.

This is an interesting finding for several reasons. First, it implies that semantic opposition is frequently expressed with antonymous concepts, not being restricted to word pairs from the same syntactic category. Because of this, Fellbaum argues that not only adjectives but also at least some nouns and verbs are organized in the mental lexicon in terms of the lexical relation of antonymy. Second, in relation to our study it seems that because cross-categorical opposites cannot be substituted for one another, such pairs are unlikely to be found by means of automatically generated patterns.

2.3.4 *Canonicity and discourse functions of opposites*

Much of the recent corpus-based research on opposites has focused on canonical opposites and their properties, mostly because the defining characteristics of canonicity, and exactly which pairs can be considered canonical and which non-canonical is not at all clear (Jones et al. [2007]). Such studies relate to the goals of automatic antonym harvesting in a limited way. Canonical opposites themselves are of little interest: the set of canonical opposites is restricted, and well-studied. The relevant aspect of these studies for the current work is that they relied on manually identified patterns that were used to study antonym canonicity. But while they used a small number of intuitive patterns for examining a number of canonical pairs they contain, we extend this idea by generating thousands of patterns automatically and examining the range of pairs of opposites they extract. Below we present previous work in detail to explain the relationship between patterns and opposites.

Jones et al. [2007] suggested that besides significant co-occurrence, the number of different patterns opposites occur in, or their "*breadth of co-occurrence*", should be used to determine which opposites are canonical. Fourteen variations of seven textual patterns from Jones's earlier work (Jones [2002]) were selected and Google was used to find patterns with pre-selected opposites filling the first or the second adjective slot: [*dull and X alike*], and [*X and dull alike*]. The variation helped to establish how recip-

rocal the antonymous relationship was to a given adjective pair. The retrieved opposites were then ranked, taking into account the number of different pattern types in which each pair occurred. This number was then compared with the frequency with which the pair occurred together. For the adjective *dull* for example, *bright* and *dynamic* both co-occurred with *dull* a comparable number of times (103 and 83 respectively) yet *bright* occurs in eleven of the fourteen patterns, while dynamic only in three, suggesting *bright* and *dull* might be a more canonical pair. This seems to be confirmed by antonym elicitation tasks, where *bright* was also the number one response for the stimulus *dull* (Paradis and Willners [2007]). This result was used to support [Jones et al., 2007]'s claim that patterns, or the range of contexts in which a pair occurs, and whether or not the pair was reciprocal, are all strong indicators of antonym canonicity.

The *breadth of co-occurrence* may be a particularly relevant feature: occurring in more than one context might contribute different information about the nature or strength of a candidate pair than frequency or co-occurrence statistics alone. But on the other hand, this suggests that a pair occurring in only one pattern, however frequent it may be, may not be a good pair. The problem with practically applying this finding to our work is that many antonymous pairs will not be frequent even in very productive patterns in a large corpus. Then the question remains whether such pairs can still be retrieved and evaluated automatically by means of textual patterns.

Using newspaper data to develop a classification of antonym usage, Jones [2002] was the first to do solid empirical work on the functions of opposites in context. His goal was to identify the different textual functions of opposites and their frequency. To this end he selected a list of 56 traditionally recognized antonym pairs including gradable and non-gradable opposites. For each pair he extracted all sentences that contained both members of the pair from a 280 million word corpus from the newspaper *The Independent* and then manually selected a sample of 3,000 sentences from the total set of 55,411 extracted sentences. This sentence set was then used to define and classify lexico-syntactic patterns in which opposites co-occurred.

Jones distinguished eight textual functions of canonical opposites, of which six were indicated by lexico-syntactic patterns. The largest textual function with reliable patterns was Coordinated Antonymy, making up 38.4% of all 3,000 examples. This function was found with patterns like [*both* <*ANT*> *and* <*ANT*>], [<*ANT*> *or* <*ANT*>], [<*ANT*> *as well as* <*ANT*>]. Opposites in these patterns are said to signal inclusiveness or exhaustiveness of scale (Jones [2002], p.61). Distinguished Antonymy, in which there is an emphasis on the distinction between the two groups, was charac-

terized by patterns like [*the difference between <ANT> and <ANT>*] and [*separating <ANT> and <ANT>*]. Other textual functions included Comparative Antonymy (patterns like [*more right than wrong*]), Transitional Antonymy (patterns like [*from success to failure*]), Negated Antonymy (patterns like [*in success, not failure*]) and Extreme Antonymy (patterns like [*to the very young and the very old*]).

Note that the most frequent (38.7%) textual function, Ancillary Antonymy, was not defined by any patterns. These were examples where an antonym pair indicated or emphasized an opposition between a pair of words or phrases that would not necessarily be in opposition otherwise. For example, a well-known antonymous pair *love-hate* in "*I love to cook but I hate doing the dishes*" (modified from Jones [2002], p.45, example 5a) is used to emphasize another opposition, namely, cooking is contrasted with washing the dishes, and the writer's affinity for both tasks emphasizes this contrast.

Recall that to extract the original sample of sentences, Jones relied on a list of opposites where both words belonged to the same syntactic category. This might imply that the resulting sample was limited in that sentences where antonymous concepts were expressed by words from different word classes were omitted from it. Even so, Ancillary Antonymy was one of the largest identified classes, suggesting that if cross-categorical pairs were also added (cf. Fellbaum [1995]), it would be the most frequent textual function. This again points out that lexico-syntactic patterns restrict the identification of opposites to the pairs of words of the same syntactic category, neglecting cross-categorical pairs that express antonymous concepts.

The categories discussed above are not necessarily exhaustive. Moreover, since most of the opposites freely occurred in several types of patterns, the patterns do not coincide with the traditional classification of opposites. Jones himself [2002: Chapter 9] notes that he did not find any relationship between the traditional categories of opposites and their textual functions.

Although these corpus-based studies give insights into the functions of opposites in discourse and their canonicity, their main shortcoming is that the pairs as well as the patterns used were identified manually, making the proportion of types unreliable. It is therefore not at all clear whether, for example, patterns of type Coordinating would be more useful than other pattern types for identifying good antonym pairs. Previous corpus work on antonymy also does not explain how new pairs become contrastive and how to identify them automatically without giving at least one member of the pair.

In summary, the main conclusion of the studies of Justeson and Katz [1991], Fellbaum [1995], Jones [2002] and Jones et al. [2007] in relation to our work is that their

findings show that a pattern-based method for finding opposites automatically is a plausible way to pursue. Automatic extraction of patterns has the advantage of not being influenced by biases and intuitions of the researcher. It is superior to manual identification because it requires less time and provides flexibility within different genres and languages, it is easier to extend, and it guarantees some measure of consistency and coverage. To our knowledge, there are no studies that aim at automatic identification of patterns for finding opposites. However, pattern-based methods have been successfully used in automatic extraction of such lexical semantic relations as hyponymy, for example, *car - vehicle* and meronymy, for example, *finger - hand*. We will now present these methods to explain the development and the current state-of-affairs of pattern-based methods in relation extraction, arguing that patterns can also be applied to finding opposites.

2.4 Corpus-driven research in relation extraction

2.4.1 Pattern-based methods in relation extraction

The original work on patterns in relation extraction was done by Hearst [1992] who suggested that patterns in which two words co-occur can signal lexical semantic relationships between them and, therefore, can be used to identify those relations. Using six manually identified surface patterns with part-of-speech information like [*such X/Noun as Y/Noun*], she found phrases like [*such authors as Shakespeare*] or [*such injuries as ulceration*] and used them to successfully extract facts such as that *Shakespeare* is a kind of *author* and *ulceration* is a kind of *injury*. In the 8.6 million corpus of encyclopaedia texts, Hearst found 153 candidate hyponym pairs, of which 61 were listed in a hyponym relationship in WordNet (Fellbaum [1998]), suggesting that the method could easily add useful relations to WordNet that were missing. As future work, Hearst suggested that a similar approach can be used to identify other lexical relationships.

Testing Hearst's suggestion, Berland and Charniak [1999] applied a textual pattern-based method to meronym extraction using a newspaper corpus of 100 million words. Starting with a set of manually chosen meronym pairs, they extracted all sentences that contained them and manually identified plausible surface part-of-speech patterns. The best two patterns, namely, [<*WHOLE*> */Noun's* <*PART*> */Noun*] as in *building's basement* and [<*PART*> */Noun of a|the* <*WHOLE*> */Noun*] as in *basement of a build-*

ing, were then enlisted to extract new meronyms for six single words used as seeds, for example, the seed word *car*. Found words were automatically scored and ranked based on several probability metrics. They report an accuracy of 55% for the top 50 meronyms derived from six seeds and an accuracy of 70% for the top 20 meronyms derived from six seeds based on the majority vote of the evaluation of the pairs by five judges.

Evaluation of the results is one of the major problems with this type of work. Berland and Charniak used five judges for evaluation of found pairs, but as the authors point out themselves: "Lacking a formal definition of part, we can only define those words as correct and the rest as wrong. While the scoring is admittedly not perfect, it provides an adequate reference result." [1999, p.60]. Unfortunately, they do not mention the agreement score between participants but they indicate that to simplify the evaluation task, their goal was to find only nouns denoting physical objects. Still they report that their participants "often disagreed" leading to a "fair consensus" only [1999, p.57].

Evaluation of pairs found by our method faces similar problems. Instead of a definition that covers all instances of opposites, it is possible to give examples of typical opposites and ask participants to use them as guidelines. The majority vote in such cases might be not perfect but the assessment of the agreement between participants can further indicate how closely participants agreed on the evaluation of found pairs. This is a widely used method of evaluation in automatic relation harvesting. Berland and Charniak also compared the top-20 parts of the word *car* with WORDNET (Fellbaum [1998]) and found that their method missed important parts (for example, *engine* and *door*) but also found many parts not listed in WORDNET (for example, *tailpipe*, *break* and *speedometer*). This shows that evaluation based on computational lexical resources like WORDNET, which were fully or partially constructed by hand, can be incomplete and misleading. And the results from automatic extraction of lexical semantic relations can improve existing computational resources used for many NLP applications. We will discuss evaluation of our results in detail in chapter 3.

Like Hearst [1992], Berland and Charniak did not automate the pattern identification step. They assumed that the two selected patterns were both frequent and precise to successfully identify the target relation. It is not clear how many meronym pairs and which ones were used to select patterns. Interestingly, Hearst [1992] also had tried her method on meronym extraction but without success, arguing that only hyponymy can be identified by means of patterns. It may be that Berland and Charniak's results

were better due to a bigger corpus, as well as a more sophisticated ranking of found pairs. However, this also highlights that manual identification of patterns based on researchers' intuitions is inconsistent and that differently chosen patterns lead to different results. To what extent manual selection of patterns can affect the results is not clear. But it is plausible to conclude that an approach in which patterns are generated automatically is more advantageous as it is faster, it is applicable to more extensive data collections of different genres and it might find patterns that would otherwise be missed.

Using the minimally supervised bootstrapping algorithm *Espresso*, Pantel and Pennacchiotti [2006] identified generic[1] surface part-of-speech patterns automatically. The patterns were then used to extract a range of lexical semantic relations like meronymy and hyponymy as well as more specific semantic relations like reaction and succession.[2] Also beginning with seed pairs, they extracted all sentences in which these pairs co-occurred in a 6.3 million words newspaper corpus and used the sentences to generalize patterns. All patterns were automatically evaluated. The scoring was calculated as an association measure between a given pattern and highly reliable instances based on pointwise mutual information (Church and Hanks [1990]). The top-10 best patterns were used to find new pairs. Extracted pairs were also evaluated using an association score between a given pair and a highly reliable pattern. Since generic patterns are frequent and they contain a lot of noise, pattern recall was increased by using the Web to retrieve more instances. They showed that in comparison to the results reported in similar studies on relation extraction, their method of using automatically extracted generic patterns had high precision but also high recall. The obtained precision scores for the sample of 50 extracted instances of hyponyms and 50 extracted instances of meronyms with their top algorithm were between 73% and 85%. These results are based on the evaluation of found pairs by two participants. They report a Kappa-score of 0.69 indicating sufficient agreement between two judges.

Results of these previous pattern-based methods have shown that automatic lexical extraction can be very fruitful. They also suggest that we should not expect to achieve precision scores above 70-80%. However, it has yet to be established how realistic this method is for automatic extraction of opposites, as it appears to be a more

[1] Generic patterns are patterns with high recall and low precision.

[2] Reaction is defined as a relation that occurs between chemical elements that can be combined in a chemical reaction. For example, *hydrogen gas* reacts with *oxygen gas*. Succession is defined as a relation that indicates that a person succeeds another person in a position or title. For example, *George Bush* succeeded *Bill Clinton*.

difficult task than hyponym or meronym extraction for several reasons. First, it is not known whether patterns can successfully deal with relations expressed by different part-of-speech categories. Second, even those patterns that have been established to indicate contrast (for example, [*the difference between <ANT> and <ANT>*]) can contain many non-opposites unlike patterns for finding meronyms (like [*<NP> is an instance of <NP>*]).

Note that none of the previous pattern-based studies have addressed the question of finding opposites. This may be due to the difficulty of identifying patterns that reliably contain opposites. It may also be that previous studies did not deal with opposites because antonymy is still a not well-defined lexical semantic relationship. Nevertheless, opposites have been consistently mentioned in the literature that describes distributional methods (see section 2.4.2) for finding synonyms, as opposites are frequently found in their results as noise. Because of that, several attempts have been done in the studies that use distributional methods to identify opposites and consequently to separate them from near-synonyms. We will now discuss studies that identified opposites using distributional methods and show that while such methods can be used to study the strength of antonymy, they cannot be used to separate opposites from other distributionally similar words. More importantly, we will argue that while distributional methods can be used to validate existing opposites and the strength of antonymy between them, pattern-based methods offer a more powerful way for finding existing as well as *novel* opposites.

2.4.2 *Distributional methods and opposites*

Distributional methods are based on the idea that the context of a word can tell about its meaning. This is also known as the Distributional Hypothesis (Harris [1954]). Some words share the same contexts and will be found together, other words share the same contexts but do not co-occur together within those contexts. Semantically *related* words like *doctor* and *hospital* tend to appear together and can be found within close proximity to each other. Semantically *similar* words like *rich* and *wealthy*, on the other hand, share many similar contexts but are unlikely to be found together. The context itself can be defined in different ways. *Co-occurrence methods* define context as the *n* number of words that surround the target word. *Syntax-based methods* identify context in terms of a syntactic relation between the target word and the second word.

Distributional methods have been widely used for finding synonyms. But one of the

well-known problems with such methods is that they also find opposites. For example, Grefenstette [1992] presented a syntax-based method for extraction of semantically similar words from raw texts based on the distance similarity measure that calculates the number of shared features, that is contexts, between two words. He argued that the number of shared contexts reflects the strength of association for two words so that the more features words share, the more strongly they are associated with each other even if they do not appear in the same sentence or the same document. Grefenstette found that in the results "[...] a great number of closest modifiers seemed to be antonyms" (p.63). In fact, his system identified 33 of Deese's 39 opposites, or 85%, as the closest or next-to-closest pairs. Since opposites are similar in all but one respect, they not only tend to co-occur with each other but they are also likely to modify the same contexts.

Mohammad et al. [2008] examined whether distributional methods could be used to measure the strength of association between opposites. Following this idea, the authors proposed an unsupervised co-occurrence method for determining what they refer to as the *degrees of antonymy* between word pairs. The degrees are meant to reflect intuitions of speakers found in psycholinguistic experiments that show that some opposites are perceived as 'better' (such as *thin* - *thick*) than others (such as *thin* - *chubby*). For each target word pair, based on the thesaurus categories, their approach first decides whether a pair is antonymous or not. If yes, based on the co-occurrence statistics, it decides whether the pair has a high or low degree of antonymy.

The degrees of antonymy are related to antonym canonicity studied by Jones et al. [2007], as opposites with more degrees are more typical and consequently canonical than opposites with less degrees. However, an important difference between Mohammad and colleagues' method and the pattern-based approach taken by Jones and colleagues is that the latter can only be applied for finding the most canonical opposite for a target word overall, while the former approach identifies the most suitable opposite for a target word within a given set of candidate pairs.

Candidate opposites were identified using automatically generated sets of contrasting seed pairs. The first seed set consisted of 2,734 word pairs derived by means of 16 morphological rules (such as *X* - *imX* for *possible* - *impossible*). In particular, each rule was applied to each word in the Macquarie thesaurus and if the resulting token was also found in the thesaurus, two words were considered a seed pair. This method generated some non-opposites, for example, the pair *sect* - *insect* in which both words are encoded in the thesaurus but are not opposites. Although the authors do not specify how many non-opposite pairs were found, they suggest that such cases were rare and

did not impact the results. But this simple heuristic on its own seems to be a productive method for identifying morphologically-related opposites and can be extended to find novel pairs in specific domains, such as medical texts.

The second seed set of 10,807 pairs was obtained from WORDNET (Fellbaum [1998]). In particular, for each pair of words that were linked in WORDNET as opposites, all words from the same synsets were matched as contrasting pairs. Note, that this method extracted 20,611 candidate pairs, 47.6% of which (9,804 pairs) were discarded as they were not found in the Macquarie thesaurus, the resource used to identify whether a word pair was antonymous or not. This already points out one of the main shortcomings of an extraction method based on available lexical resources, which is their limited coverage of the target relation. As a consequence, such method restricts the range of found *novel* pairs and instead it validates already *established* pairs. A pattern-based method, on the other hand, does not face this constraint.

Once the seeds were generated, for each seed pair, if $word_1$ was found in the Macquarie thesaurus's category C_1 and $word_2$ was found in the category C_2, the two categories were identified as contrasting. Then, if two candidate words belonged to two categories identified as contrasting, they were classified as opposites. When the words occurred within the same thesaurus paragraph, they were considered to have high degree of antonymy. Otherwise, the degree of antonymy was estimated using distributional metrics based on pointwise mutual information (PMI, Church and Hanks [1990]) used to determine how likely two words were to co-occur together in the text. The co-occurrence statistics was obtained from the *Google n-gram corpus* (Brants and Franz [2006]).

In addition the authors also examined a heuristic for identifying opposites, according to which all adjacent categories in the thesaurus were treated as contrasting.

The system was evaluated on a set of 950 closest-opposite questions, which consisted of a target word and five alternatives with more than one possible opposite of different degrees of antonymy. For example, for the target word *adulterate*, the system correctly identified *purify* as its closest opposite, although another opposite *correct* was also present. When the system could not find an opposite for the target word, it discarded the question. The system achieved the best precision score of 0.83 using adjacency heuristics only, although half of the questions were discarded leading to the recall of 0.46. The best overall performance was obtained when both seeds and the adjacency heuristic were used, achieving a precision score of 0.76 and a recall score of 0.64.

The fact that only 20 pairs (0.2%) in the question set directly matched one of the 10,807 seeds from the WORDNET illustrates the need to have a method for finding opposites that is not constrained by any lexical resource, such as pattern-based methods presented in this dissertation.

Given that the antonym identification step was based on the thesaurus, it is not clear whether the method was more efficient for identification of opposites expressed by a particular part-of-speech category, as it might contain more categories for noun pairs as opposed to adjectives and verbs.

2.4.2.1 "Distributional Hypothesis of Antonyms"

Mohammad and colleagues' method is based on what the authors call the *Distributional Hypothesis of Antonyms* which is a synthesis of the *Co-occurrence Hypothesis* proposed by Charles and Miller [1989] and the *Substitutability Hypothesis* proposed by Justeson and Katz [1991]. A valuable contribution of their work is that the authors provide statistical evidence to support both hypotheses and suggest that significant co-occurrence as well as shared contexts can be used as useful cues to determine opposites. Although the authors use co-occurrence statistics to determine degrees of antonymy rather than to filter out non-opposites from the results, their findings have a direct impact on any study of opposites that aims at identifying novel pairs. That is why we will discuss this part of their work in detail.

To examine the *Co-occurrence Hypothesis*, two sets of 1,000 pairs each consisting of adjective - adjective, noun - noun and verb - verb pairs were randomly selected from the WordNet. The first set contained opposites while the second set was used as a control set that consisted of unrelated words. First, Mohammad et al. [2008] counted the number of times each word occurred individually and the number of times both words co-occurred in the same sentence in a window of five words in the British National Corpus (BNC) (Burnard [2000]). They then calculated the mutual information for each of the word pairs and averaged it. The average mutual information between the words in the set of opposites was 0.94 with a standard deviation of 2.27. The average mutual information between the words in the control set was 0.01 with a standard deviation of 0.37. Thus, antonymous word pairs co-occurred together significantly more often than chance ($p<0.01$). What this means is that we can use significant co-occurrence to identify and discard unwanted non-opposites from our automatically found pairs. This was not done in Mohammad et al. [2008]'s study because, as the authors point out

themselves, significant co-occurrence is not sufficient for identifying *only* opposites as also collocations tend to co-occur more often than chance. However, while not sufficient, using significant co-occurrence can substantially reduce noise in the results and using this strategy is a plausible solution that we employ in our experiments.

To examine the *Substitutability Hypothesis*, the same sets of word pairs were used to test whether opposites occur in similar contexts more often than non-opposites. Using distributional measures of distance based on the pointwise mutual information, Mohammad et al. [2008] calculated the distributional distance between each of the senses for each word pair. They then averaged the distance between the closest senses of the word pairs for all pairs in each set. On the scale from 0 (unrelated) to 1 (identical) the control set had an average semantic closeness of 0.23 with a standard deviation of 0.11 while the antonymous word pairs had an average semantic closeness of 0.30 with a standard deviation of 0.23. This means that, in comparison to other pairs, opposites tend to occur in similar contexts ($p<0.01$). As the authors mention themselves, shared contexts are not sufficient for identifying only opposites as also near-synonyms tend to occur in similar contexts, that is, contexts in which they can be substituted for one another. This highlights an important limitation of the study of Mohammad et al. [2008], namely, that although they show that a distributional method can be used to find the *closest* opposite among a given set of pairs, it is not a suitable approach for finding *novel* pairs of opposites.

2.4.2.2 Limitations of distributional methods for finding opposites

Grefenstette's system would not be able to separate near-synonyms from opposites. A pattern-based method, on the other hand, can deal with this problem as it is based on the sentential co-occurrence of words rather than their substitutability for one another in similar contexts and consequently it does not find near-synonyms. This has been discussed in the study of Lin et al. [2003] whose goal was to automatically identify and filter out opposites from their results on automatic acquisition of distributionally similar words. They proposed to use two textual patterns, referred to as *patterns of incompatibility*, namely [*from X to Y*] and [*either X or Y*], to calculate how often a given pair occurred with one of the two patterns on the Web. Their assumption was that words that appear in these patterns are very likely to be semantically incompatible and therefore they cannot be synonymous. As a test, they searched for co-occurrences of the synonyms *adversary* - *opponent* and opposites *adversary* - *ally* using the Web.

They found that out of the 2,469 hits returned by AltaVista for the query with *adversary - ally*, this pair co-occurred in the selected patterns 30 times (1.2%). The same query with *adversary - opponent* returned 2,797 hits, but the pair was not found with either of the patterns. To evaluate their method, Lin and colleagues computed distributional similarity between 45,000 words from a newspaper corpus and randomly selected 80 pairs of synonyms and 80 pairs of opposites that were among the top-50 distributionally similar words in the results and that were also present in Webster's Collegiate Thesaurus (Kay [1988]). Their pattern-based method for determining whether a pair was a synonym or not achieved an 86.4% precision and a 95% recall.

Since their evaluation set consisted of synonyms and opposites only, it is difficult to interpret the results in respect to the usefulness of the patterns of incompatibility for identification of opposites as it is not clear whether such patterns would also contain other relations (for example, co-hyponyms like *apple, pear, orange*). However, their results point out that methods based on distributional similarity alone cannot separate synonyms from opposites, while a pattern-based method that uses just two rather general patterns of incompatibility can.

Turney [2008] extended [Lin et al., 2003]'s idea in their work. They proposed a supervised machine learning algorithm for identification of several lexical semantic relations including near-synonyms like *levied - imposed* and opposites like *black - white*. As input, the algorithm took a training set of word pairs with class labels and a testing set of word pairs without labels. Each word pair was represented as a vector in a feature space and a supervised learning algorithm was used to classify the feature vectors. The elements in the feature vectors were based on the frequencies of automatically generated textual patterns taken from a large corpus of web pages (about 280 GB of plain text). The output of the algorithm was an assignment of labels to the word pairs in the testing set. For disambiguation between near-synonyms and opposites, the system was tested on a set of 136 English as a second language (ESL) practice questions. Using ten-fold cross-validation, the system achieved an accuracy of 75%. However, always guessing the majority class resulted in an accuracy of 65.4%. As a conclusion, the authors suggest that the strength of their approach is not its performance on any particular task, but the range of the tasks it could handle.

Summing up, pattern-based methods for finding opposites have many advantages over distributional methods. First, as has been already mentioned, patterns find co-occurring pairs so they are unlikely to find near-synonyms. Second, a pattern-based method seems to be more appropriate for applying to corpora which are not sense-

annotated. While [Mohammad et al., 2008]' results suggest that pairs of opposites co-occur significantly often regardless of their intended sense (p.985), this is not enough to identify that *hot* - *cold* are opposites in the case *hot* is used in its sense of TEMPERATURE, for example, *hot water* but not in its sense of SPICE, for example, *hot curry*). It is not possible to annotate word senses in the large corpora currently used for automatic relation extraction as it is too expensive to be done manually and automatic systems for word sense disambiguation perform poorly. To deal with this problem, Mohammad et al. [2008] use categories defined in the Macquarie thesaurus. In particular, using distributional distance between two thesaurus categories, they consider two words to be antonymous in the senses from those categories that are closest to each other. For example, *play* is antonymous to *work* only in its sense of ACTIVITY FOR FUN and not DRAMA. They find that the thesaurus category containing *work* is closer to the category containing *play* in the sense of activity for fun and those senses are selected as contrasting. The problem with such an approach is that it is directly related to the number of categories covered by the thesaurus. Recall that out of 20,611 contrastive pairs derived from WORDNET, only 10,807 of them were found in the Macquarie thesaurus and could be used in the study of Mohammad et al. [2008]. Further recall that when one of the words from a found pair was not present in the target category, the pair was discarded as not antonymous, again eliminating potential opposites. As a result, such an approach cannot be used to identify novel pairs of opposites not covered by existing lexical resources. A pattern-based method can tackle this problem from a different angle. In particular, it has been shown that in cohesive texts words that co-occur together tend to be close in meaning (Halliday and Hasan [1976]) suggesting that when opposites co-occur with each other within a sentence, they are likely to be used with senses that refer to the same category. Similarly, the findings of Justeson and Katz (Justeson and Katz [1991]) and later Fellbaum (Fellbaum [1995]), who argues that antonymous concepts rather than words tend to co-occur together, also suggest that opposites found by means of patterns are likely to have contrasting rather than unrelated senses. Since patterns are likely to contain words with related senses, a pattern-based method is not dependent on any lexical resource. This points out yet another advantage of such methods - their independence from the available computational resources, making it easier to extend such work to languages other than English. Interestingly, while previous studies agree that antonymy is not well covered in the WORDNET and many useful opposites remain uncovered in this resource, it remains to be the main source used in the experiments for identification and validation of good opposites. A solution to

this would be a method that does not depend on the lexical resources for finding novel instances of opposites. This is exactly what a pattern-based method can offer.

CHAPTER 3

Means of evaluation of found pairs

The performance of any relation extraction algorithm must be evaluated. The goal of this chapter is to introduce two evaluation methods commonly used in relation extraction and to give a full understanding as to how and why we chose certain evaluation methods to classify pairs found by means of the proposed algorithms in the studies presented in Chapter 4, Chapter 5, and Chapter 6.

Section 3.1.1 describes evaluation based on computational lexical resources, in particular WORDNET (Fellbaum [1998]), which is widely used for English, and CORNETTO (Horak et al. [2008]), which is widely used for Dutch. This section also describes previous studies that relied on these resources and outlines the main advantages and limitations of this method. Section 3.1.2 presents an evaluation method based on manual classification of found pairs. We explain in detail how inter-annotator agreement can be assessed, introducing the notion of Kappa-scores. This is followed by a discussion on how precision scores are estimated once all pairs are classified by participants. This section also describes the shortcomings of manual classification.

We argue that because each previously proposed evaluation method on its own has its shortcomings and advantages, the best way of evaluating found pairs is by using a combination of these evaluation methods.

3.1 Evaluation methods used in relation extraction

Although one of the main goals of finding lexical semantic relations automatically is to improve and augment the coverage of these relations in existing computational lexical resources such as WORDNET (Fellbaum [1998]), lexical resources continue to be one of the most widespread means for the evaluation of automatically found results. In addition, evaluation is often based on the participants who evaluate all found pairs or a sample of randomly selected pairs. Finally, evaluation can also be done by means of existing corpus-based dictionaries, such as COBUILD (Sinclair [2003]) and manually annotated datasets specifically created to serve as gold standards. As will be discussed below, each of these methods has its advantages but also limitations. For this reason we will use a combination of lexical resources and manual evaluation to assess the quality of our results.

3.1.1 Computational lexical resources

Hearst [1992], who was the first to use hand-crafted textual patterns like [<*Noun*> *is an instance of* <*Noun*>]) to find hyponym-hypernym pairs like *car - vehicle*, used WORDNET to evaluate her results. WORDNET (Fellbaum [1998]) is a manually constructed computational lexical resource in which word senses are organized into *synsets*, that is, unordered sets of near-synonyms. For example, one of the synsets for the word *car* contains *car*, *auto*, *automobile*, *machine*, *motorcar*. The synsets are hierarchically structured by the hyponym-hypernym relation. For example, the synset with *car* is under the synset *vehicle*, the synset with *vehicle* is under the synset *transport* and so on. Membership in multiple synsets reflects that a given word has more than one meaning (that it is polysemous). Because this resource consists of pairs and relations manually classified by trained experts, all annotations are reliable. This is one of the main advantages of using hand-crafted lexical resources for evaluation of automatically extracted relations. Another advantage is that such evaluation can be done fast and a large number of extracted pairs can be evaluated at once. The process is cheap as it does not require additional expenses, such as training of annotators in the case of manual evaluation.

Going back to Hearst [1992], she examined how many found noun - noun hyponym - hypernym candidate pairs were present in WORDNET's hierarchy, and how many of them were in the hyponym-hypernym relation. Among 152 found pairs, 226 were

unique words, 180 of which were present in WORDNET and only 33.9% (61 words) of which were connected by the hyponymy relation. It is not clear whether all found pairs were also evaluated by judges but Hearst used her results to argue that pattern-based automatic extraction of lexical relations is useful for improving the coverage of the hyponymy relation in the lexical resource, implicitly suggesting that all 180 found pairs that were present in WORDNET could be connected with each other as hyponym-hypernyms. Her findings directly point out the main limitation of the evaluation of candidate pairs by means of computational lexical resources such as WORDNET. Namely, that such resources are limited by the number of pairs manually, or semi-manually, captured in the resource. As a result, using them for evaluation of novel found pairs can be misleading as the relationship between good novel pairs might be missing in the resource.

This limitation has a direct impact on evaluation of candidate opposites in the studies presented in this dissertation in that the coverage of antonymy in the existing lexical databases is smaller than that of hyponymy and synonymy. Thus, even more pairs might simply be missing in such resources. Given that hyponymy is the main relation for the organization of synsets in WORDNET, its coverage of hyponym-hypernym pairs is rather substantial, providing a source for the evaluation of found pairs in the studies on hyponymy extraction. But there is no available comparative analysis as to the limitations of using lexical resources such as WORDNET for evaluation of the results of different lexical semantic relations. To give an idea of how useful WORDNET can be in relation extraction we will now discuss a few studies in which this resource was used for evaluation of meronyms and synonyms. In those studies, in addition to WORDNET, the authors also used judges to evaluate their results. Consequently, they used manual judgements as gold standards in evaluation of the coverage of the target relation in WORDNET in direct relation to their results.

Berland and Charniak [1999] relied on WORDNET for the evaluation of automatically found meronyms, also known as part-of relation, for example, a *petal* is part of a *rose*. Instead of evaluating all pairs, they examined subsets of pairs, in particular, the top-20 highest ranked words for seed words. Their results showed that WORDNET both contained and missed good pairs. For example, for the seed word *car*, ten words listed as parts in WORDNET were missing from the top 20 candidate meronyms (including *car* paired with *door*, *engine*, and *gear*). On the other hand 16 of 20 automatically found candidate meronyms were missing from WORDNET, including *car* paired with *radiator*, *break*, *bumper* and others. Such inconsistencies were prevalent even for com-

mon word pairs like meronyms *car - bumper*. Similar to the findings of Hearst [1992], the authors concluded that their results could be used to expand WORDNET. As a more informative method for the evaluation of their results, Berland and Charniak used five judges. For each seed word (like *car*), participants were asked to rate a set of 100 words, of which 50 were automatically found candidate meronyms. Participants were unaware of the goal of the experiment. Out of the top 20 meronym candidates for the word *car*, 17 were marked by the participants as correct. This result illustrates that manual evaluation can be a more indicative way for estimating the performance of a lexical extraction algorithm as it covers *all* pairs found by a proposed algorithm. A more detailed discussion of manual evaluation is presented in Section 3.1.2.

In relation to our work, Berland and Charniak's results provide evidence to support the usefulness of using not only computational resources but also manual evaluation for a more balanced and indicative evaluation of the performance of a relation extraction algorithm, in our case for finding opposites.

Note that earlier studies on automatic extraction of lexical relations aimed at finding candidate pairs that would further be evaluated by annotators. This was possible due to a rather limited number of pairs they found. More recent studies, however, try to reduce manual intervention as much as possible. Despite its limitations, WORDNET has become the primary resource used as the gold standard for evaluation of results as well as for training of classifiers (for example, work of Snow et al. [2005], and Snow et al. [2006]).

Work on relation extraction in Dutch uses either the Dutch part of EuroWORDNET (available since 1999) or a more recently available computational lexical semantic resource for Dutch CORNETTO (Horak et al. [2008]) (available since 2008). **CORNETTO** is based on two existing databases for Dutch: the Dutch part of EuroWordNet (Vossen et al. [1999]) and the Referentie Bestand Nederlands (Maks et al. [1999]). Similar to the original WORDNET (Fellbaum [1998]), word senses in CORNETTO are organized into *synsets*, which are hierarchically structured by hyponym-hypernym relations. The database contains approximately 70k synsets, which altogether contain over 91k lemmas (70k nouns, 9k verbs, 12k adjectives) corresponding to 118k word senses. It also encodes meronymy and antonymy. Antonymy relation is reported between 1,588 word senses, or over 5k opposites.

Unfortunately, in comparison to the original WORDNET, Dutch resources are smaller and contain less lexical information. The study of Hofmann and Tjong Kim Sang [2007], who replicated work of Snow et al. [2005] for finding hyponym-hypernym pairs

in Dutch, illustrates the limitation of using WORDNET-like resources like CORNETTO especially for languages other than English. Similarly to the study in English, which relied on English WORDNET, Hofmann and Tjong Kim Sang [2007] used all nouns covered in the Dutch WORDNET to train their classifiers. The best performance scores were nevertheless lower than the scores reported for English in Snow et al. [2005]. The authors suggested that the main reason for the difference in the results was the discrepancy in the overall number of nouns covered in the English WORDNET (namely, 116,648 nouns) and the Dutch part of EuroWORDNET (namely, 45,981 nouns). This is a clear illustration that lexical extraction based on existing lexical resources is severely constrained by the coverage of manually constructed databases even for hyponymy. This limitation is particularly severe for languages other than English, because in such cases the databases are either smaller than their English counterparts or they are not available at all.

Fortunately, computational lexical resources for Dutch have been steadily improving. Nevertheless, even studies on synonyms, the defining relation in WORDNET as synsets are at the core of the organization of Dutch WORDNET, show that such resources are insufficient for evaluation of automatically found synonyms. In particular, van der Plas and Bouma [2005], report that 60% of synonym candidate pairs returned by their system as most similar words to a list of 1,000 test words taken from Dutch WORDNET were not found in this computational lexical resource. Even when both words were present, the relations between them were often missing. In a similar vein, van der Plas and Tiedemann [2006], who also present experiments on finding synonyms, report that 37% of pairs found by their system and judged as synonyms by judges, had both words present in Dutch WORDNET but not linked as synonyms.

In summary, using WORDNET and its deviants like CORNETTO for evaluation of automatically found pairs has several advantages but also a few constraints. Its main advantage is that such databases are manually constructed by experts. As a result, all annotations are reliable. This implies that using CORNETTO instead of judges does not require any costs in terms of time and expenses it takes to train the experts and in terms of time it takes to evaluate pairs. Finally, it allows a large number of pairs to be evaluated at once. Among its flaws, the main limitation of the resource-based evaluation is that the coverage of word pairs and relations is constrained and incomplete even for well-studied relations. Especially domain-specific words might be lacking in the resources. For example, adjectives *chronic* and *acute* are opposites only in the medical domain and would not be recognized as such outside of the medical context.

This might be problematic also in the evaluation of our results because since we use a corpus of newspaper texts with a lot of domain-specific terms that might be missing in the currently available lexical resources.

Also, manually updating a resource is a costly process, difficult to implement. Because of this, many novel pairs are missing, especially, for specific domains and languages other than English.

As has been discussed above, the problems related to the resource-based evaluation pose similar problems for studies that aim at finding different lexical semantic relations. The extent to which they are affected might depend on the type of the relation. But even evaluation of automatically found synonyms and hyponyms, two relations prevalent such computational lexical resources as WORDNET, are affected by the shortcomings discussed above.

It is also unknown what the differences in regard to the evaluation of relations expressed by different part-of-speech categories are. All studies discussed in this section were occupied with noun - noun pairs. Studies on extraction of opposites presented in this dissertation are occupied with antonymy expressed by adjectives, nouns and verbs. It is possible that the extent of usefulness of a resource-based evaluation will depend on the syntactic category of found pairs.

Taking all aforementioned points into account, we used CORNETTO as the first step of evaluation for pairs found in the studies presented in this dissertation. We chose CORNETTO, as it is the largest available resource for Dutch and it includes relations covered in Dutch WORDNET. Unfortunately, antonymy is not fully covered in this resource, especially for opposites expressed by nouns and verbs. That is why, in addition to CORNETTO, we compiled a list of opposites from an online dictionary *Mijnwoordenboek.nl*. This dictionary contains a total of 1,228 unique antonym pairs, only 271 of which (22%) are also present in CORNETTO. Note that we did not use the widely-known *Van Dale* dictionary for evaluation of the results because the *Van Dale* Dutch-English dictionary (Martin and Tops [1989]), the *Van Dale* English-Dutch dictionary (Martin and Tops [1986]), as well as a lexical database provided by *Van Dale* were all used as a base for the Dutch WordNet, which is included in CORNETTO.

In contrast to studies discussed above, evaluation of automatically found opposites is a particularly challenging task because many good opposites are simply missing from any of the available resources. When opposites are covered, their representation varies drastically across different databases even for the most typical opposites. For example, according to the LONGMAN *Dictionary of Contemporary English* [2003], the word *fat*

has only one opposite: *thin*. In WORDNET, *fat* has two direct opposites: *thin* and *nonfat* and two indirect opposites: *unprofitable* and *unfruitful*[1]. In *Collins* COBUILD *Advanced Learner's English Dictionary* [2003] the opposites of *fat* are *thin* and *slim*. None of the resources list such opposites of *fat* as *skinny*. Paradis and Willners [2007] also note that even when opposites are covered in a dictionary, they are often represented asymmetrically so that only 37% of opposites listed in the COBUILD dictionary are given in both directions. That is, while *dead - alive* and *alive - dead* are both listed as opposites, *new* is listed as an opposite of *old* but not the other way around.

With this in mind, in addition to CORNETTO and *Mijnwoordenboek.nl*, manual evaluation was used as a second step of evaluation of found candidate pairs. The exact procedure is presented in the next section.

3.1.2 Evaluation of the results by participants

Since the coverage of the existing lexical resources is limited and because many studies use corpora from a specific domain (like newspaper texts or Wikipedia texts), resources like WORDNET might be insufficient for evaluation of the results. For example, recall that out of top 20 candidate meronyms for the word *car*, automatically found in the study of Berland and Charniak [1999], 17 candidates were identified by judges as meronyms, whereas 16 candidates were not even present in WORDNET. For these reasons, many authors prefer to use manual evaluation instead of lexical resources. In such cases, pairs are usually classified as either belonging to the category of interest or not. For example, in the study of Ittoo and Bouma [2010], who aimed at finding meronyms, two participants annotated all found pairs as meronyms and non-meronyms. Such classification can be further used to calculate the *precision scores*. All pairs classified as meronyms unanimously, that is, by both judges, are treated as *true positives*, all pairs unanimously judged as non-opposites are treated as *false positives* and pairs that do not receive unanimous votes, that is, one judge classifies a pair as meronyms while the other judge classifies it as non-meronyms, are usually discarded. The precision score is then computed as follows:

[1] WORDNET distinguishes between direct, that is, lexical, and indirect, that is, conceptual, opposites. In the former case, the pairs consists of two conventional opposites, for example, *rich - poor*, *hot - cold*. In the latter case, opposition is mediated through synonymy, for example, the antonymy of *rich* and *destitute* is mediated by the similarity of *destitute* to *poor*.

$$Precision = \frac{N_{truepositives}}{N_{truepositives} + N_{falsepositives}}$$

To evaluate found candidate opposites, we use the same method for calculating the precision scores but instead of two judges we employ three. All three participants brought in for evaluation of pairs presented in this dissertation were students at the Faculty of Natural Sciences at the University of Groningen. In a 'Yes/No' classification task, each pair of found words, whose automatic scoring was higher than a set threshold, was presented on the screen and participants had to classify it as opposites or non-opposites. The participants could go back and change their answer for any pair as many times as they wanted. There were also no time constraints on the completion of the task. The evaluation was implemented as a Python program that participants could run on their own computer at any convenient time. The pairs were divided into equal text files, where each file contained approximately 200 pairs. Once the evaluation task started, the participants had to complete the evaluation of all pairs in a file. The participants were paid eight Euro per hour, the time it took them to complete the task was measured automatically.

The main limitation of the manual evaluation of any relation extraction algorithm is that participants often disagree with each other as to the categories they assign to the pairs. Partially, this is because very often it is difficult to come up with a simple yet clear definition of a target relation that covers most well-established instances of the relation as well as non-typical pairs.

For example, Berland and Charniak [1999] mentioned that the scores for meronyms based on manual evaluation differed greatly among different seed words because there was no formal definition of parts or linguistic tests to enable participants to unambiguously distinguish between parts and non-parts. As an example, the authors mention that everyone recognized that a *shifter* is part of a *car* but not everyone accepted *production* as part of a *plant*. Such judgements are very much based on participants' intuitions, leading to a low level of agreement among participants. As a results, their judgements might become unreliable.

Opposites are notoriously difficult to define, so instead of providing participants with a strict definition of antonymy, we presented them with a number of typical opposites expressed by different part-of-speech categories at the beginning of each session. In addition, the participants completed a training session with immediate feedback, in which they had to recognize opposites from a set of well-established opposites and

randomly selected non-opposites.

One of the requirements for manual annotation to be valid is that all annotators need to be consistent with each other when evaluating the results. In other words, it is important that participants consistently annotate the same pairs. To determine reliability of annotation, participants must evaluate the same pairs using the same guidelines, preferably consisting of a clear definition of each category. Given that the above requirements are fulfilled, it is then possible to calculate *inter-annotator agreement* using *Kappa-score*. This score reflects whether participants consistently made the same judgements or whether they disagreed with each other.

Different measurements are used to assess inter-annotator agreement. For example, when only two participants are evaluating pairs, each pair will be judged either as a true positive (correct) or as a false positive (incorrect). Once more than two participants are used, each pair can be judged as a true positive by all three participants, receiving *unanimous vote*, it can also be judged by all participants as a false positive, again receiving *unanimous vote*, or it can be judged as a true positive only by the majority of participants but not all of them, in this case, receiving *majority vote*. For example, in case we have three participants, if the pair *city - countryside* is judged as opposites by all three participants, it receives unanimous vote as a true positive. The pair can also be judged by all three participants as non-opposites. In this case it will receive unanimous vote as a false positive. But it can also be that two participants classify it as opposites while one participant classifies it as non-opposites or that two participants classify it as non-opposites and one participant classifies it as opposites. In the latter cases, the pair is said to receive majority vote.

Reliability of agreement between *more than two participants* is usually calculated by means of the *Fleiss's kappa score* (Fleiss [1971]). Fleiss's kappa score is defined as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\bar{P} - \bar{P}_e$ is the degree of agreement achieved above chance (based on observed level of agreement for each annotation of each case for each participants and the extent to which participants agreed for each case), and $1 - \bar{P}_e$ is the degree of agreement that is achieved above chance.

As an illustration, let's calculate Fleiss's kappa score for four pairs (n = 4) that were classified by three participants (K = 3) according to two categories (m = 2) (see Table

Found word pairs	Total number of judgements Category 1	Category 2
word pair 1	0	3
word pair 2	1	2
word pair 3	3	0
word pair 4	3	0
Total	*7*	*5*

Table 3.1: An illustration of classification tableau for four pairs (n = 4), judged by three participants (K = 3) as belonging to one of the two categories (m = 2) using a binary answer yes/no (1 or 0).

3.1).

$\bar{P}$ represents observed level of agreement, it is measured in the following way:

$$\bar{P} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}^2 - nK}{nK(K-1)}$$

In our example, $\bar{P}$ equals to 0.83 as can be seen below:

$$\bar{P} = \frac{((0^2+3^2)+(1^2+2^2)+(3^2+0^2)+(3^2+0^2))-4*3}{4*3(3-1)} = 0.83$$

Estimated agreement $\bar{P}_e$ is the sum of square probability of each category (P_j) on the whole dataset:

$$\bar{P}_e = \sum_j P_j^2$$

where

$$P_j = \frac{1}{nK}\sum_{n=1}^{i} K_{ij}$$

In our example, participants chose in total Category 1 seven times (0 + 1 + 3 + 3) and Category 2 five times (3 + 2 + 0 + 0). P_j is then equal to 0.51 and the Fleiss's kappa score is 0.65:

$$P_j = (\frac{7}{12})^2(\frac{5}{12})^2 = \frac{74}{144} = 0.51$$

$$\kappa = \frac{0.83 - 0.51}{1 - 0.5138} = 0.65$$

When participants completely agree with each other on the whole set of candidate pairs, the κ score is equal to 1. If there is no agreement whatsoever, the κ score is equal to 0. There are many suggestions as to how to interpret kappa scores between 0 and 1. For interpretation of the results for Dutch, we will use the scale originally proposed in Landis and Koch [1977] as a rule of thumb. A kappa score of $\leq$0.2 will be considered to indicate poor agreement, a kappa score between $>$0.2$\leq$0.4 will be considered to indicate fair agreement, a kappa score between $>$0.4$\leq$0.6 will be considered to indicate moderate agreement, a kappa score of $>$0.6$\leq$0.8 will be considered to indicate strong agreement, and a kappa score of $>$0.8 will be considered to indicate a near perfect agreement.

Using levels of agreement we can conclude that for the example above participants strongly agreed with each other.

Note that the reason why there are different scales for interpreting kappa scores is because some tasks are intrinsically easier than others and, as a result, they receive higher kappa scores. For example, kappa scores for part-of-speech tagging, syntactic annotations and similar tasks go as high as 0.98, whereas kappa scores for annotation of lexical semantic relations, discourse annotations and other evaluation tasks that often involve subjective interpretations are usually between 0.65 and 0.7. This is because it is much harder to give a clear definition of a lexical relation than to define syntactic categories that can be objectively identified by (trained) judges.

In summary, manual evaluation can be used to classify automatically found candidate opposites. To ensure that classification is reliable, it is necessary to calculate inter-annotator agreement. Inter-annotator agreement scores for evaluation of lexical semantic relations are lower than those reported for other NLP tasks partially due to the difficulty of defining the target relation in a way that makes it possible for participants to make consistent judgements. For the optimal evaluation of our results, it is therefore useful to use a combination of evaluation steps that involve computational lexical resources as well as judgements.

CHAPTER 4

Performance of textual patterns for finding opposites

Corpus-based studies on opposites suggest that when they co-occur within a sentence, opposites often appear in so-called ***surface textual patterns***, or constructions like [*the difference between <ANT> and <ANT>*] or [*from <ANT> to <ANT>*]. So far such patterns have been manually identified to manually analyse and classify functions of opposites in discourse (Jones [2002]) and to study their canonicity (Jones et al. [2007]). In this chapter, we go a step further and examine whether textual patterns can be used to *automatically* identify opposites in large corpora. Unlike previous work, our textual patterns are acquired and scored automatically. We show that such patterns are more specific than manually identified ones but many of them can be generalized to match existing pattern types discussed in relation to opposition. We also find patterns that do not match any hand-crafted patterns, suggesting that automatic extraction of patterns provides a more consistent and reliable method for finding productive patterns that would be missed otherwise. This supports similar findings in studies on automatic meronym and hyponym extraction (Berland and Charniak [1999], Ittoo and Bouma [2010]).

4.1 Inspirations for the present study

Work presented in this chapter was inspired by previous research in several research areas. In the first part of this section we outline previous studies, describing how their findings are related to the current work.[1] In the second part of the section we present our expectations for the results based on previous work.

4.1.1 Where it all began: antonymy in psycholinguistics experiments

In relation to the research done on antonymy, we drew our inspiration predominantly from *corpus-based studies* whose main focus was analyses of opposites, their types and their behaviour in sentences found in real data. First of all, our work is based on the assumption that opposites co-occur with each other within a sentence significantly more often than would be expected by chance. This is known as the *Co-occurrence Hypothesis*, which was originally proposed by Charles and Miller [1989], who wanted to explain a high association between adjectival opposites found in psycholinguistic experiments (Deese [1964]). Two ideas in this hypothesis are relevant to us. First, the fact that opposites can be found in pairs within a sentence implies that there is no need to use larger chunks of text to find pairs of opposites automatically. Second, it means that we can use significant co-occurrence as a means of identifying and separating relevant pairs (that is, candidate opposites) from noise in our automatically found results by taking into consideration only those pairs that co-occurred with each other significantly more often than expected by chance.

4.1.2 Patterns and antonymy: original corpus-based studies

There are several reasons why we decided to automatically find opposites using a pattern-based method. First, corpus-based studies that immediately followed Charles and Miller's work suggest that opposites co-occur in patterns. In particular, Justeson and Katz [1991] examined the frequencies of intersentential occurrences of adjectival opposites in a one million word Brown Corpus (Burnard [2000]). They found that the adjectival opposites they studied co-occurred with each other in the corpus significantly more often than would be expected by chance. They also found that in many sentences

[1] Chapter 2 provides a thorough discussion of literature that is related to the current work in full detail including studies outlined in this section.

adjectival opposites co-occurred with each other in specific constructions like [*between* <*ANT*> *and* <*ANT*>]. In relation to our work, their findings suggest that adjectival opposites can be automatically identified by means of constructions that we will refer to as *surface textual patterns*.

So far we described studies that only looked at adjective - adjective opposites. Using the Brown Corpus (Burnard [2000]), Fellbaum [1995] conducted a first large scale study, examining sentential co-occurrence of opposites expressed by adjectives, nouns and verbs. The author found that all opposites, regardless of their syntactic category, co-occurred with each other in the corpus significantly more often than would be expected by chance. However, she found that only opposites expressed by adjectives co-occurred in textual patterns. This would mean that although sentential significant co-occurrence could be used to identify opposites from non-opposites, a pattern-based method would not be able to deal with the extraction of opposites expressed by nouns and verbs.

The most recent corpus-based work, however, suggests the opposite. In particular, Jones [2002], Jones et al. [2007], Paradis and Willners [2007], Paradis et al. [2009] argue that opposites expressed by all syntactic categories co-occur sententially in textual patterns. We assume that the size of the corpus plays a role and that opposites expressed by part-of-speech categories other than adjectives will be found in corpora of larger size. A larger corpus is necessary for finding noun - noun and verb - verb opposites in patterns as all aforementioned studies used much larger corpora (including Google search) than Fellbaum [1995]. What this means for our study is that our algorithm will be able to find opposites expressed by all three syntactic categories we examine (namely, adjectives, nouns and verbs) given that we use a very large corpus of approximately 450 million words. However, to examine whether our assumption about the size of the corpus is correct, it is necessary to test our algorithm on corpora of different sizes. This will tell us whether or not corpora of larger size are indeed necessary for finding opposites expressed by nouns and verbs. Given that previous studies suggest that opposites expressed by adjectives can be identified in corpora that would be considered small, it seems that opposites expressed by nouns and verbs require larger data collections.

4.1.3 *Computational linguistics: existing work on pattern-based methods for relation extraction*

Existing computational work on automatic extraction of pairs of hyponym - hypernyms, for example, *car - vehicle*, meronyms, for example, *petal - rose* and other lexical-semantic relations provides overwhelming evidence to support the usage of a pattern-based method for finding opposites (particularly, Hearst [1992], Pantel and Pennacchiotti [2006], Snow et al. [2005], Snow et al. [2006] for hyponymy extraction in English and Tjong Kim Sang and Hofmann [2007], Hofmann and Tjong Kim Sang [2007], Tjong Kim Sang and Hofmann [2009] for hyponymy extraction in Dutch; Berland and Charniak [1999], Pantel and Pennacchiotti [2006] for meronym extraction in English and Ittoo and Bouma [2010] for meronym extraction in Dutch)[1].

First and foremost, the original work of Hearst [1992] showed that a small number of hand-crafted patterns like [<*W1*> *is a kind of* <*W2*>], six in total, could be used to automatically identify hyponym-hypernym pairs like *Shakespeare - author*. Hearst argued that patterns in which two words co-occur can signal not only hyponymy but various lexical semantic relationships between the two words, suggesting that such patterns can be used to identify those relations. Note that Hearst was successful at finding hyponym - hypernyms expressed by nouns, using an 8.6 million word corpus of encyclopaedia texts. Since finding opposites seems to require much larger corpora, in relation to our work Hearst's results suggest that patterns that contain hyponym-hypernym noun - noun pairs are qualitatively different from patterns that contain opposites (at least, for noun - noun pairs). In particular, textual patterns that contain opposites might be more general and more noisy than textual patterns that contain hyponym - hypernyms. As a result, hyponym-hypernym noun - noun pairs are found in patterns in much smaller corpora than patterns that contain antonymous noun - noun pairs. Nevertheless, above all, Hearst's results show that a pattern-based method can be used to find lexical-semantic relations.

4.1.4 *Automatic extraction of textual patterns: why this is a necessary step*

Note that in our work textual patterns used for finding candidate opposites are generated automatically. Pantel and Pennacchiotti [2006] were the first to show that patterns

[1] An interested reader is referred to Chapter 2 for a detailed discussion of these studies.

for automatic identification of hyponymy, meronymy and a few more specific relations like succession can be generated automatically. To do that they used sets of seed pairs, that is well-known agreed upon examples, to find all sentences in which these pairs co-occurred in a 6.3 million word newspaper corpus. Patterns were then automatically generated from found sentences. Because some of the best patterns, according to the automatic scoring, were too general, for example a pattern [<*W1*> *and* <*W2*>], Pantel and Pennacchiotti used the Web as an additional tool to improve the recall of such patterns. This step improved the performance of their algorithm leading to both high precision and high recall. Their findings provide evidence that pattern-based methods can be applied to successfully identify various lexical semantic relations at least between noun - noun pairs and that the patterns themselves can be identified automatically as well.

In contrast to manual selection, automatic identification of patterns has multiple advantages. First, generating patterns automatically is fast. It requires less time than a researcher would take to go through every single sentence by hand. As a result, it is applicable to extensive data collections of different genres. It might find patterns that would otherwise be missed by a participant based on his/her intuition alone. There are no studies that examine the extent to which manual selection of patterns can affect the results. However, it seems that a researcher's intuition can be misleading and good productive patterns can be missed. For example, Hearst [1992] used her approach with manually selected patterns to find not only hyponyms but also meronyms. Her patterns were successful at finding hyponyms but patterns for finding meronyms did not yield good results. This led Hearst to conclude that a pattern-based method might not work for meronyms. However, this is not the case, as Berland and Charniak [1999], who had a similar algorithm based on other manually identified patterns, were able to find meronyms. This difference in the results shows that different researchers give preference to different hand-crafted patterns, which leads to large inconsistencies in the results. To ensure that we overcome flaws of manual pattern selection, all our patterns are generated automatically.

4.1.5 Corpus genre and size matter

Interestingly, another difference between the study of Hearst [1992] and Berland and Charniak [1999] is the genre and the size of the corpus they used. Namely, Hearst used a collection of encyclopaedia texts of 8.6 million words whereas Berland and

Charniak used a newspaper corpus of 100 million words. We assume that the genre of the corpus plays a role as newspaper texts can contain a wider range of patterns while encyclopaedia texts tend to have a more repetitive structure. To examine whether the genre of the corpus plays a role on antonym extraction, we also ran our experiment on a corpus of encyclopaedia texts.

In relation to the size of the corpus, recall that Fellbaum [1995] was able to find opposites using just a one million words corpus. This suggests that we do not necessarily need a vast amount of data to find good opposites automatically. However, the results might differ per part-of-speech category of the seed set. To test whether there are difference in relation to the size of the corpus, we conducted the same experiments on subcorpora of smaller sizes. In particular, we created a subcorpus with the first 100 million words and a subcorpus with the first 200 million words from the full corpus.

4.1.6 Looking beyond antonym canonicity

A final point that we would like to discuss is the relation between our work and corpus-based work on antonym canonicity, originally proposed by Jones and colleagues [2006] who argue that textual patterns can be used to determine antonym canonicity. In particular, the authors suggest that the number of different *types* of patterns in which opposites occur is indicative of their canonicity. For example, the pair *rich - poor* is canonical because these opposites co-occur with each other in more than ten different pattern types. The pair *wealthy - poor* is not canonical because these opposites co-occur with each other in fewer than three pattern types. Jones et al. [2007] refer to the number of pattern types, in which a pair of opposites co-occurs, as their *breadth of co-occurrence*, arguing that this is one of the two major factors for some opposites to be perceived as "better", or more canonical, than others. In our study, we can examine whether there is a difference between well-established canonical opposites and non-canonical opposites that we find in respect to the number of automatically-generated patterns in which they are found. This will allow us to overcome one of the main shortcomings of Jones and colleagues' study, namely, the fact that they did not study the *breadth of co-occurrence* for non-adjectival opposites. This will give us a more exact picture as to the actual behaviour of opposites in the corpus.

4.2 Assumptions

Based on the previous findings described above, we have the following assumptions from the results of our pattern-based algorithm for finding opposites:

1. **Automatic identification of opposites**:
   - only significantly co-occurring pairs can be treated as candidate opposites;
   - opposites found automatically will be expressed by all three part-of-speech categories;
   - well-established canonical opposites will be found in a wider range of automatically identified pattern types than non-canonical opposites.
2. **Automatic identification of patterns**:
   - given a large enough corpus, it is possible to identify useful surface textual patterns automatically;
   - automatically generated textual patterns can successfully find good opposites.
3. **Size of the corpus**:
   - more noun - noun and verb - verb opposites will be found in larger corpora;
   - noun - noun and verb - verb seeds will find more opposites in larger corpora.
4. **Genre of the corpus**:
   - newspaper texts will lead to the extraction of a larger number of pattern types than encyclopaedia texts.

4.3 Method

This section gives an overview of the methodology used. The details on the corpora used are presented in Section 4.3.1. The seed sets are discussed in Section 4.3.2. The algorithm is presented in 4.3.3, followed by a detailed explanation of automatic scoring of generated patterns and found pairs in Section 4.3.4.

4.3.1 Corpora

We used the Twente Nieuws Corpus (TwNC, Ordelman [2002]). This corpus of written Dutch consists of newspaper texts and subtitle texts covering three years (1999-2002) of publications. It is made up of approximately 450 million words. The version of the corpus we used was preprocessed by the Alpino parser (van Noord [2006]). The corpus was tokenized, that is punctuation marks were separated from words and sentence boundaries were identified, and lemmatized, that is all words were reduced to their base forms.

To examine whether similar results can be achieved on a smaller corpus, we created two subsets of the corpus, one contained the first 100 million words of the TwNC and second one contained the first 200 million words. The results obtained from the subcorpora were compared with the complete corpus.

The same experiments were also conducted using a collection of Dutch Wikipedia[1] texts. This corpus consisted of approximately 127 million words and 6.8 million sentences. This corpus was tokenized but not lemmatized. Wikipedia[2] is a web-based encyclopaedia written and edited by volunteers around the world in various languages.

4.3.2 Seeds

Three sets of seeds of different sizes were compiled for each of the three part-of-speech categories: adjectives, nouns and verbs. The sets consisted of opposites whose canonicity has been established in earlier studies by word association tests (Deese [1964]), in corpus analysis (Jones et al. [2007]) or taken from theoretical classifications of opposites (Cruse [1986]). A preliminary pilot study showed that these seeds performed better than automatically extracted morphologically-related pairs (for example, *known* - ***un****known*). The sets consisted of six, 12 and 18 seed pairs where the set of 12 seeds included the six seeds and six additional pairs and the set of 18 seeds included the 12 seeds and six additional pairs. The summary of all pairs is presented in Table 4.1.

[1] http://nl.wikipedia.org/wiki/Hoofdpagina
[2] www.wikipedia.org

Nr. of seeds	Adjective Dutch	seeds English	Noun Dutch	seeds English	Verb Dutch	seeds English
	arm - rijk	poor - rich	begin - eind	beginning - end	verlies - win	lose - win
	open - dicht	open - closed	man - vrouw	man - woman	geef - neem	give - take
6	groot - klein	large - small	dag - nacht	day - night	koop - verkoop	buy - sell
	snel - langzaam	fast - slow	vraag - antwoord	question - answer	open - sluit	open - close
	mooi - lelijk	beautiful - ugly	voordeel - nadeel	advantage - disadvantage	vind - verlies	find - lose
	nauw - breed	narrow - broad	vrede - oorlog	peace - war	lach - huil	laugh - cry
	droog - nat	dry - wet	top - bodem	top - bottom	eindig - begin	end - begin
	nieuw - oud	new - old	hemel - hel	heaven - hell	stijg - daal	increase - decrease
12	hoog - laag	high - low	uitgang - ingang	exit - entrance	spaar - besteed	save - spend
	koud - warm	cold - hot	sterkte - zwakte	strength - weakness	bevestig - ontken	confirm - deny
	oud - jong	old - young	straf - beloning	punishment - reward	slaag - misluk	succeed - fail
	lang - kort	long - short	optimist - pessimist	optimist - pessimist	vraag - beantwoord	ask - answer
	blij - verdrietig	happy - sad	echtgenoot - echtgenote	husband - wife	val aan - verdedig	attack - defend
	actief - passief	active - passive	chaos - orde	chaos - order	haat - houd van	hate - love
18	goed - fout	right - wrong	roofdier - prooi	predator - prey	daal - neem toe	fall - rise
	dood - levend	dead - alive	werkgever - werknemer	employer - employee	sluit uit - voeg toe	exclude - include
	zwaar - licht	heavy - light	feit - fictie	fact - fiction	exporteer - importeer	export - import
	hard - zacht	hard - soft	aanval - verdediging	attack- defense	voeg toe - verwijder	add - remove

Table 4.1: List of seed pairs for each part-of-speech category. The 12 seed set includes the first six seeds plus six additional antonym pairs. The 18 seed set includes the set of 12 seeds plus six additional opposites.

4.3.3 Algorithm

The algorithm can be divided into five steps. First, all sentences that contained both halves of any of the predefined seed pairs were extracted from the corpus (**Step 1**). Each of these sentences was used to generate *all* possible textual patterns of consecutive words given a minimum and maximum pattern length (**Step 2**). We set the minimum length at three tokens and the maximum length at seven tokens. With shorter sentences, the length of a sentence was the maximum length of patterns. Patterns could consist of words as well as punctuation marks and numerals. By allowing a large difference between the minimum and the maximum length, we can examine whether productive patterns are more likely to be shorter and more general, for example, patterns like [*<ANT> and <ANT>*], or longer and more specific, for example, patterns like [*a transition from <ANT> to <ANT> and*].

Seeds in found textual patterns were substituted by <-1>, to be used as place-holders. To illustrate generation of patterns, consider the example below:

(1) "Prada is een mix van **rijk** en **arm**, van verschillende culturen, van vroeger en nu", aldus Miuccia.
"Prada is a mix of **rich** and **poor**, of different cultures, of the past and the present", according to Miuccia.

Sentence (1) contains seed opposites *arm - rijk* “poor - rich” that occur at a distance of one token from each other. This means that a total of 15 patterns with a length ranging from three to seven elements containing both opposites can be generated from this sentence (see Table 4.2)[1].

Obtained patterns do not contain any information about part-of-speech categories of the found instances of seed pairs and the possible pairs in the place-holders. That is, we do not disambiguate between an occurrence of *arm* “poor” and *rijk* “rich” as an adjective - adjective pair as in “*the difference between rich and poor countries*” and an occurrence of *arm* “poor” and *rijk* “rich” as a noun - noun pair as in “*The gap between the rich and the poor widens*”. Sentences containing the pairs regardless of the part-of-speech category would be extracted and used to generate and evaluate patterns.

There are also sentences that contain both halves of a seed pair but at a distance greater than seven tokens. For example, it is not possible to generate a pattern for *rich* -

[1]Note that the algorithm was running on the version of the corpus containing digits instead of words, numbers and punctuation marks to improve the efficiency of the algorithm. Examples of patterns are presented with words for illustration purposes.

Nr.	Dutch	English
1	<-1> en <-1>	<-1> and <-1>
2	van <-1> en <-1>	of <-1> and <-1>
3	<-1> en <-1> ,	<-1> and <-1> ,
4	van <-1> en <-1> ,	of <-1> and <-1> ,
5	<-1> en <-1> , van	<-1> and <-1> , of
6	mix van <-1> en <-1>	mix of <-1> and <-1>
7	<-1> en <-1> , van verschil	<-1> and <-1> , of different
8	mix van <-1> en <-1> ,	mix of <-1> and <-1> ,
9	een mix van <-1> en <-1>	a mix of <-1> and <-1>
10	van <-1> en <-1> , van	of <-1> and <-1> , of
11	is een mix van <-1> en <-1>	is a mix of <-1> and <-1>
12	een mix van <-1> en <-1> ,	a mix of <-1> and <-1> ,
13	mix van <-1> en <-1> , van	mix of <-1> and <-1> , of
14	van <-1> en <-1> , van verschil	of <-1> and <-1> , of different
15	<-1> en <-1> , van verschil cultuur	<-1> and <-1> , of different cultures

Table 4.2: All possible textual patterns acquired from sentence (1).

poor from a sentence like "*The rich countries must open their borders for the products from poor countries*". Such sentences were discarded because patterns of this length would be too specific.

Finally, there were sentences that contained one or both halves of a seed pair more than once, as in a sentence "*Poor people in rich countries have little in common with poor people in poor countries*". Such instances were very rare and the preference was given to the opposite words that were closest to each other. Thus, in the sentence above, patterns were generated for the first occurrence of the word *poor* with *rich* in the range of [<*poor*> *people in* <*rich*> *countries have little*].

Once all patterns were generated, the corpus was searched for all occurrences of the patterns where the positions of the wildcard tokens "<-1>" could be taken by any word (**Step 3**). Patterns that were found only once were eliminated. The rest of the patterns were automatically scored. Patterns with a score lower than a set threshold τ were dismissed (**Step 4**). Finally, based on the scoring of patterns, candidate word pairs that filled the wildcard positions were also automatically scored and sorted in the descending order (**Step 5**). Extracted pairs that contained numerals, punctuation marks and frequent words from the stop list were removed (**Step 6**).

A detailed description of the automatic scoring of patterns and pairs is discussed next.

4.3.4 *Automatic scoring of patterns*

All generated patterns that occurred more than once and contained at least one of the seed pairs were automatically scored. This scoring was later used to automatically evaluate and rank extracted candidate pairs. A straightforward way to compute how likely it is for a pattern$_i$ to contain an antonym pair would be by estimating its **conditional probability** as defined below:

$$SCond(pattern_i) = P(Rel_{ant}|Pat_i) = \frac{F_i}{N_i} \tag{4.1}$$

where F$_i$ is the number of times pattern$_i$ contained one of the seed pairs and N$_i$ is the number of times pattern$_i$ was found overall. It is the probability that pattern$_i$ contains the relationship of antonymy (Rel$_{ant}$).Intuitively, patterns that extract a lot of seeds receive higher scores. However, using direct counting of pattern frequencies to estimate probabilities can be tricky and this particular evaluation has two direct shortcomings.

The first is related to the fact that if the corpus used contains a very infrequent pattern that was found only with the seeds, this pattern will be amongst the patterns with the highest scores. More text would reveal instances of this pattern in which it does not contain seeds but this is impossible to know without more data.

Another issue with this scoring metric is that no difference is assumed between frequent and infrequent patterns as long as they contain the same *proportion* of seed pairs. Consider the following example as an illustration: pattern$_A$ was found twice with two seed pairs, pattern$_B$ was found 50 times and in all occurrences it contained seed pairs. According to SCond($pattern_i$), both patterns have the same absolute score: SCond(P_A) = 2/2 = 1 and SCond(P_B) = 50/50 = 1. Although theoretically both patterns are equally likely to contain antonyms, it seems reasonable to treat a pattern that was found more often to be more reliable than a pattern that was found only twice. Further, consider another pair of patterns, one of which occurred 25 times and contained seeds twice and the other occurred 250 times and contained seeds 20 times. Again, in both cases the scoring will be the same: SCond(P_C) = 2/25 = 0.08 and SCond(P_D) = 20/250 = 0.08. That is, both patterns have the same eight percent probability of containing an antonym pair. However, although theoretically there should be no difference between the two, it seems plausible that in practice patterns that occur more frequently might be more reliable than a pattern that occurs rarely.

To deal with infrequent patterns, one can use the *add-one smoothing* operation originally proposed by Church [1988] which presumes that the data contains an extra pos-

itive and an extra negative (unseen) pair for each pattern (4.2):

$$S\_AddOne(pattern_i) = P\_AddOne(Rel_{ant}|Pat_i) = \frac{F_i + 1}{N_i + 2} \tag{4.2}$$

If the original score of pattern$_A$ was derived as 2/2, that is the pattern contained seeds in all of its occurrences (all positive pairs), we add one additional positive pair to the two that were found and one positive and one negative pair to the total number of pattern occurrences. Thus, S_AddOne(P_A) = (2 + 1)/(2 + 2) = 0.75. Note that with add-one smoothing more frequent patterns have higher scores, so S_AddOne(P_B) = (50 + 1)/(50 + 2) = 0.98.

Unfortunately, add-one smoothing treats frequent and infrequent patterns that contain the same proportion of seed opposites as equally good (that is, productive) patterns. Moreover, when two patterns have proportionally the same number of seeds, add-one smoothing favours less frequent patterns: S_AddOne(P_C) = 3/27 = 0.11 while S_AddOne(P_D) = 21/252 = 0.084 and S_AddOne(P_E) = 201/2502 = 0.08. To take this difference into account we followed an approach similar to Riloff [1996] and Thelen and Riloff [2002], who instead of add-one smoothing, modified the scoring function by computing the logarithm of the total number of times pattern$_i$ occurred:

$$S\_RlogF(pattern_i) = P\_RlogF(Rel_{ant}|Pat_i) = \frac{F_i}{N_i} \times \log_2(F_i) \tag{4.3}$$

The logarithm reduces the influence of very infrequent patterns, favouring patterns that extract a high number of seeds and patterns that occur often and contain a moderate number of seeds. So, although pattern$_C$, which occurs 25 times and contains seeds twice, has the same proportion of seeds as pattern$_D$, which occurs 250 times and contains seeds 20 times, the scoring of pattern$_C$ will be lower than the scoring of pattern$_D$: S_RlogF(P_C) = $(2/25) \times \log_2(2) = 0.08$ and S_RlogF(P_D) = $(20/250) \times \log_2(20) = 0.35$.

While this scoring function captures the differences between frequent and infrequent patterns by giving higher scores to more frequent ones, it is difficult to interpret the scoring any further. That is, while conditional probability scores gave percentage estimations for each pattern (as an eight percent chance of containing an antonym pair in the example above), this scoring can only be interpreted in relation to the patterns and the relative scores between them, so pattern$_D$ is better than pattern$_C$ but we cannot say anything further as to how probable it is for pattern$_D$ to contain antonyms. To correct for this, we modified the formula as follows:

$$S\_pattern_i = \sin(\frac{F_i \times \frac{\pi}{2}}{N_i + c}) \tag{4.4}$$

where c was a small constant to prevent the denominator of the above formula to be zero. After preliminary testing, the value of c was set to 5. This scoring function favors patterns that contain the largest number of seeds and patterns that are most frequent:

S_P_A = $\sin(\frac{2\times\frac{\pi}{2}}{2+5}) = 0.43$;

S_P_B = $\sin(\frac{50\times\frac{\pi}{2}}{50+5}) = 0.98$;

S_P_C = $\sin(\frac{2\times\frac{\pi}{2}}{25+5}) = 0.1$;

S_P_D = $\sin(\frac{20\times\frac{\pi}{2}}{250+5}) = 0.12$.

This scoring method has the same preferences as S_RlogF(pattern$_i$) but in addition it is possible to interpret the scoring as probabilities. So, pattern$_B$ has the highest probability of containing antonym pairs (98%), followed by pattern$_A$ (43%), pattern$_D$ (12%), and finally pattern$_C$ with the probability of 10%. Using this scoring method, all patterns that occurred more than once were scored. Patterns with a scoring lower than the threshold τ set to 0.1 were discarded.

4.3.5 Automatic scoring of pairs

Using the scores of the remaining patterns, we calculated the 'antonymy score' for each new instance that was found in conjunction with these patterns. The scoring was based on the number of times a pair was found in each pattern and patterns' scoring:

$$AntS(pair_j) = 1 - \prod_j (1 - S(P_i))^{C_{ij}} \tag{4.5}$$

where S(P$_i$) is the score of pattern$_i$ and C_{ij} is how often the j-th pair occurred in the i-th pattern. Consider an example with pair$_x$ that was found once with pattern$_A$ and twice with pattern$_D$. If the pattern scores are given by S_(P_A) = 0.43 and S_(P_D) = 0.12, then the antonymy score can be calculated as follows:

AntS(pair$_x$) = $1 - (1 - 0.43)^1 \times (1 - 0.12)^2 = 0.56$

Intuitively, this is the probability that it is *not* the case that all evidence for the pair being an antonym pair is false. In the example above, pair$_x$ is 56% likely to be antonymous. At the end, all pairs were ranked according to their scores. Pairs with

the scoring $\geq$0.9 were evaluated by judges. Pairs that co-occurred less than five times were discarded.

4.4 Results for the corpus of newspaper texts - TwNC

In this part the results obtained from the Twente Nieuws Corpus of Dutch (TwNC) will be presented in detail. The results will be presented separately for seeds expressed by adjectives (4.4.1), nouns (4.4.2), and verbs (4.4.3). We will show that this method works and textual patterns can be used for finding good opposites as well as pairs of co-hyponyms that are contrasted with each other. We will also discuss textual patterns themselves, automatically found by the algorithm, and show that this method is capable of identifying many useful patterns which, although specific, can be classified according to the pattern types suggested by Jones [2002]. While a handful of manual patterns can limit the range of found pairs, automatic methods can identify thousands of patterns and their specificity does not restrict the results. This is an important finding as it shows not only that automatic acquisition of patterns is more productive than manual selection but the analysis of found patterns can be used to study contexts in which opposites co-occur. In other words, we can learn what kinds of word pairs can occur in patterns identified by reliable canonical seed pairs. This can also be used to study which found patterns find non-opposites.

All found pairs are discussed in relation to several means of evaluation. We discuss how many of found pairs co-occurred with each other within the corpus significantly more often than would be expected by chance to examine the extent to which the *Co-occurrence Hypothesis* (Charles and Miller [1989]) holds. This can indicate whether significant co-occurrence can be used in the future as an additional parameter for distinguishing opposites from non-opposites. This also offers an opportunity to know the range of pairs with significant co-occurrence that are found in (contrastive) textual patterns but are not opposites. In addition, found pairs will be evaluated by means of CORNETTO and *Mijnwoordenboek.nl*. We will show that these resources lack many good opposites found automatically by our method, arguing that one of the direct applications for our results is improvement of the coverage of opposites in the aforementioned lexical resources. The downside of this is that lexical resources at the moment cannot be used to evaluate candidate pairs. We will show that the best way of evaluating results is manual classification.

Finally, we present results of the same method applied to the smaller parts of the

Scoring	Pairs found with		Overlap between
	6 seeds	18 seeds	6 & 18 seeds
≥0.9	35.4% (178)	46% (483)	100% (178)
≥0.8<0.9	19.7% (99)	20.7% (217)	100% (99)
≥0.7<0.8	21.3% (107)	18.1% (190)	100% (107)
≥0.6<0.7	16.3% (82)	11.2% (117)	100% (82)
≤0.6	7.3% (37)	4% (42)	100% (37)
Total	*503*	*1,049*	*503*

Table 4.3: Total number of unique pairs found with six and 18 adjective - adjective seeds in a full version of TwNC per scoring level and the number of pairs found in both sets.

same corpus with seed sets of different sizes and show that the best results are achieved with the largest corpus using the largest number of seeds. In fact, our results suggest that the largest seed set with the smallest corpus gives similar results to the smallest seed set with the largest corpus. This is a relevant finding given that previous studies focus more on the types of seed pairs rather than their number. We discuss why the size of the seed set influences results for antonym harvesting.

4.4.1 Results for adjective - adjective seed pairs

Using a full version of TwNC, a set of six seeds extracted 503 unique pairs and a set of 18 seeds extracted 1,049 unique pairs. Table 4.3 gives an overview of the number of pairs found with each seed set at every score level. As can be seen, a smaller seed set found fewer pairs and, as is shown in column *Overlap*, all of them were also found with the set of 18 seeds. This is an interesting finding as it shows that six seeds is enough to identify 52% of pairs (546 pairs) found with 18 seeds. More than that, they receive equally high scores, although more seed pairs can influence the scoring of patterns (as more seeds give more evidence as to the productiveness of patterns) and, consequently, find different candidate pairs but there is an overlap at each scoring level. Since results found with the 18 seeds set include pairs found with the set of six seeds, pairs found with the largest set of 18 seeds will be presented first. Pairs with scoring <0.6 were discarded as their probability to be opposites was too low.

Scoring	Number of pairs	Significant co-occurrence
≥0.9	483	98.3% (475)
≥0.8<0.9	217	97.7% (212)
≥0.7<0.8	190	95.3% (181)
≥0.6<0.7	117	96.6% (113)
Total	*1,007*	97.4% (981)

Table 4.4: Number of unique pairs found with 18 adjective - adjective seeds in a full version of TwNC per scoring level, number of pairs that co-occurred with each other sententially significantly more often than would be expected by chance.

4.4.1.1 Pairs found with 18 adjective - adjective seeds

A total of 1,007 unique pairs with scoring ≥0.6 were found by the algorithm, 48% (483 pairs) of which had a scoring ≥0.9. Recall that according to Charles and Miller [1989], one of the prerequisites for two words to be antonymous, is for them to co-occur with each other within a sentence significantly more often than would be expected by chance. The overview of extracted pairs and their co-occurrences is presented in Table 4.4. As can be seen, more than 95% of found pairs co-occurred sententially in the newspaper corpus significantly more often than would be expected by chance. More pairs co-occurred significantly more often than would be expected by chance at higher score levels, reaching 98.3% (475) for pairs with the score above ≥0.9. Manual examination showed that many pairs at lower score levels with significant co-occurrence were *not* opposites, confirming earlier claims that significant co-occurrence is a necessary but not sufficient requirement for antonymy. Discarding pairs which did not co-occur significantly often improved the results (see Table 4.7 for details). Therefore, only pairs with significant co-occurrence will be discussed in the remainder of this section.

Since evaluation of found pairs in meronym and hyponym extraction is commonly done by comparing the results with manually constructed computational resources like the WORDNET (Fellbaum [1998]), the next question we address is how many found pairs at each scoring level were listed as opposites in the CORNETTO database for Dutch, which is the most similar resource to the WordNet. The results are summarised in Table 4.5. It contains score levels (column one), total number of automatically identified pairs with significant co-occurrence (column two), number of pairs listed as opposites in CORNETTO (column three) and *MWB* (column four), and the number of *unique* pairs that were opposites in one or both of the resources (column five).

Out of the total 981 pairs with scoring ≥0.6, 72.6% (712 pairs) had both words

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.9	475	17.4% (61/351)	19.2% (91)	21.7% (103)
≥0.8<0.9	212	12.7% (19/150)	10% (21)	15% (32)
≥0.7<0.8	181	8% (11/138)	5% (9)	7.2% (13)
≥0.6<0.7	113	5.5% (4/73)	5.3% (6)	8% (9)
Total	*981*	13.3% (95/712)	13% (127)	16% (157)

Table 4.5: Unique pairs found with 18 adjective - adjective seeds in TwNC significantly often per scoring level and the number of pairs that were found in one or both of the lexical resources (CORNETTO and *Mijnwoordenboek.nl* (*MWB*)).

present in this lexical resource. But only 95 of them (13.3%) were linked as opposites. More pairs listed as opposites in CORNETTO were found at higher scoring levels, going from 5.5% (four pairs) with scoring between ≥0.6 and <0.7 to 17.4% (61 pairs) with scoring ≥0.9. Among opposites with scoring ≥0.6<0.7 were pairs *civiel - militair* "civil - military", *groot - licht* "large - light", *particulier - publiek* "private - public". Among 11 found pairs with scoring between 0.7 and 0.8 and listed as opposites in CORNETTO were pairs *doorgaan - stoppen* "to continue - to stop", *jong - volwassen* "young - adult", *zuiver - onzuiver* "pure - impure". Nineteen pairs with scoring between 0.8 and 0.9 which were opposites in CORNETTO included pairs like *leeg - vol* "empty - full", *klein - lang* "small - long", *extern - intern* "external - internal". The largest number of opposites with scoring ≥0.9 (61 pairs or 17.4%) included pairs like *dun - dik* "thick - thin", *donker - licht* "dark - light", *bruto - netto* "brutto - netto", *echt - vals* "real - fake" and others.

Twenty-one pairs were linked as opposites in CORNETTO asymmetrically. For example, while *donker* "dark" was among antonym candidates of *licht* "light", *licht* "light" was not among antonym candidates of *donker* "dark". Similarly, *mannelijk* "male" was among opposites of *vrouwelijk* "female" but *vrouwelijk* "female" was not among antonym candidates of *mannelijk* "male". Other asymmetric pairs included opposites *leven - dood* "alive - dead", *dom - slim* "stupid - smart", *publiek - privaat* "public - private" and so on. Among symmetrical pairs were opposites *ernstig - licht* "serious - light", *oud - vers* "old - fresh", *oud - modern* "old - modern". This asymmetry indicates inconsistencies in the encoding of lexical information in CORNETTO and it does not reflect any underlying strength of antonymy or their canonicity.

More pairs were identified as opposites according to *MWB* than CORNETTO, par-

ticularly for pairs with the highest scoring of ≥0.9 (19.2% or 91 pairs).[1] While 65 pairs were opposites according to both resources, another 62 pairs were opposites only in *MWB* and another 30 pairs were opposites only in CORNETTO. There was no clear or systematic distinction between opposites listed in one but not the other resource, suggesting that these are simple omissions. For example, among pairs listed only in *MWB* were opposites *mooi - lelijk* "beautiful - ugly", *armoede - rijkdom* "poverty - wealth". Among pairs found only in CORNETTO were opposites *mooi - slecht* "good - bad", *groot - licht* "large - light", *actief - inactief* "active - inactive". Among pairs listed as opposites in both resources were pairs *klassiek - modern* "classical - modern", *vertrouwd - vreemd* "familiar - unusual".

Comparison of opposites listed in one of the resources but not the other allows us to analyse inconsistencies between them. For example, we can find out what kind of pairs are treated as opposites in *MWB* but are not represented in CORNETTO and the other way around. The first difference between the two is that while CORNETTO contains more pairs of opposites overall, a larger number of found pairs were listed in *MWB* than CORNETTO. Most of pairs found only in CORNETTO were adjective - adjective pairs like *civiel - militair* "civil - military", *goed - mis* "good - wrong", *los - vast* "loose - fixed", whereas the majority of pairs found only in *MWB* were noun-noun pairs like *dag - nacht* "day - night", *vijand - vriend* "enemy - friend", *amateur - prof* "amateur - professional" and so on. Thus, differences might be due to the fact that while one resource has more opposites expressed by adjectives, the other resource contains mostly noun-noun pairs.

If we look at the kinds of pairs classified as opposites in CORNETTO, we find canonical opposites expressed by adjectives (like *breed - nauw* "wide - narrow") as well as non-canonical pairs like *dochter - zoon* "daughter - son", *noorden - zuiden* "northern - southern" and even context-dependent opposites like *goed - laag* "good - low" (for example, in relation to the salary), *groot - jong* "big - young" (in relation to the age), and *groot - licht* "big - light" (in relation to risks taken). The latter might not be accepted as opposites by some theoretical approaches because they are opposites only in specific contexts but it is interesting to see that all types of these pairs are found by our approach and they are present in CORNETTO. Among pairs found only in *MWB* are canonical opposites like *heet - koud* "hot - cold", non-canonical opposites like *grof - fijn* "coarse - fine" (about salt), *gevoel - verstand* "sentiment - sense" as well as pairs

[1] It was not possible to obtain all words covered in this dictionary to know how many of found pairs had both words present in *MWB*.

like *katholiek - protestant* “Catholic - Protestant”, *cultuur - natuur* "culture - nature", and *gevolg - oorzaak* “effect - cause”. Again, the oppositeness of such pairs is context-dependent and some approaches would treat these pairs as co-hyponyms. Further in the section, we will analyse whether canonical and non-canonical pairs had different automatic scoring and whether they were found in different types of patterns.

There were also differences between resources in that they contained different antonym candidates for the same words. For example, the adjective *vrolijk* “cheerful” has one opposite in CORNETTO, namely *triest* “sad”, and another opposite in *MWB*, namely *somber*. Similarly, *zoet* “sweet” has one opposite in CORNETTO (*zout* “salty”) and three in *MWB* (*zout/zuur/bitter* “salty/sour/bitter”). These inconsistencies can be improved by extending current coverage of opposites automatically.

It is apparent from the above examples that using more than one resource for evaluation of pairs is a more reliable way to assess the results.

Note that although all seeds were expressed only by adjectives, identified textual patterns found not only adjective - adjective (*leeg - vol* “empty - full”) but also noun - noun (*aarde - hemel* “earth - heaven”) and verb - verb pairs (*doorgaan - stoppen* “to go on - to stop”). This suggests that acquired textual patterns are general enough to occur with pairs that belong to different part-of-speech categories. The best textual patterns identified by means of adjective - adjective seeds will be presented in detail in Section 4.4.1.2.

In summary, even with two resources, only 16% of pairs are identified as opposites. Given that they contain inconsistencies such as discussed above, lexical resources are not sufficient for evaluation of the results as are missing too many good opposites. Manual evaluation of found pairs can provide additional evidence as to the quality of found pairs. In addition, manual evaluation can be used to estimate the completeness of representation of opposites in CORNETTO by establishing for how many pairs judged as opposites by participants, both words are present in this resource but not listed as opposites. Manual evaluation of the results will be presented next. Since most of identified opposites were found among pairs with the scoring $\geq$0.9, only those pairs were evaluated by the participants.

All pairs were evaluated by three participants, all university students, native speakers of Dutch. In a "Yes/No" classification task, they were presented with pairs (one pair at a time) on a computer screen and their task was to decide whether a pair consisted of opposites or not. Pairs could be categorized as opposites or non-opposites by the majority vote (if two or all three participants assigned a pair to a given category). The

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	48.1% (128)	75.8% (97)	51.9% (138)	74% (102)	266
≥0.96<0.98	44.4% (24)	79.2% (19)	55.6% (30)	76.7% (23)	54
≥0.94<0.96	31.3% (20)	70% (14)	68.7% (44)	77.3% (34)	64
≥0.90<0.94	39.6% (36)	77.8% (28)	60.4% (55)	72.7% (40)	91
Total	*43.8% (208)*		*56.2% (267)*		*475*

Table 4.6: Percentage of pairs with scoring ≥0.9 extracted with 18 adjective - adjective seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

results are summarized in Table 4.6.

Participants achieved a Fleiss's kappa score of 0.66 which indicates sufficient agreement between participants, which shows a strong level of agreement for this type of task. The results show that 43.8% of found pairs (208 pairs) were judged as opposites, out of which 76% (158 pairs) received unanimous votes, that is they were judged as opposites by all three participants. Among pairs unanimously classified as opposites were pairs *gekookt - rauw* "cooked - raw", *kind - ouder* "child - parent", *knap - lelijk* "pretty - ugly", *leugen - waarheid* "lie - truth" and others. About 56% of pairs (267 pairs) were judged as non-opposites, 74.5% of those pairs received unanimous votes. Among unanimously judged non-opposites were collocations, for example, *kosten - moeite* "to cost / costs - inconvenience", *Zaterdag - Zondag* "Saturday - Sunday", correlates like *klein - lief* "small - sweet", *geel - rood* "yellow - red", and also pairs that often express opposition, especially, in the newspaper texts, for example, *Arabisch - westers* "Arabic - western", *Midden- - Oost-europees* "Central- - Eastern-European", *migrant - Nederlander* "migrant - Dutchman". Most of them were found by means of textual patterns in contrastive contexts. Therefore, such pairs can be useful for identification of Contrast relationships. Also pairs that evoked a scale but did not refer to the endpoints were unanimously judged as non-opposites. Such cases included *klein - middel-groot* "small - middle-", *hoog - midden* "tall/high - average", *lang - middel-lange* "long - middle-long". Also a pair *groen - rijp* "green - ripe" which is antonymous in the context of matureness was unanimously classified as non-opposites. This shows that non-typical opposites are not recognized by the participants as antonymous outside of the context.

This also suggests that participants have strong intuitions about opposites, preferring binary opposites, which are often mutually exclusive (one can be either a *parent* or a *child*, *pretty* or *ugly*, *raw* or *cooked*, etc.) and discarding context-dependent oppo-

Scoring level	All found pairs	Precision	Pairs with significant co-oc.	Precision
≥0.98	267	0.48	266	0.49
≥0.96 <0.98	55	0.44	54	0.45
≥0.94 <0.96	66	0.28	64	0.29
≥0.90 <0.94	95	0.39	91	0.41

Table 4.7: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 adjective - adjective seeds.

sites like *Arabic - western* and *migrant - Dutchman* which require additional context to grasp the contrast between them.

Pairs that did not receive unanimous votes as opposites or non-opposites are of particular interest to study as these groups consist of pairs that were difficult to classify. Some of the pairs judged as opposites only by the majority vote were typical opposites, for example, *dik - dun* "thick - thin", suggesting that annotators made occasional mistakes in classification. Such cases were rare. More difficult pairs for classification included, for example, kinship relationships. For example, both *dochter - moeder* "daughter - mother" and *vader - zoon* "father - son" were judged by the majority vote as opposites whereas the pair *ouder - puber* "parent - youngster" was judged as non-opposites. The difference between the two seems to be in the presence of the dimension of gender. Similarly, pairs *dier - mens* "animal - person", *dier - plant* "animal - plant", and *machine - mens* "machine - person" were judged as opposites whereas pairs *ding - mens* "object - person" and *god - mens* "god - person" were judged as non-opposites. These examples show that non-opposites by the majority vote are more similar to the opposites than unanimous non-opposites and that the distinction between them can be obscure.

Pairs *huidig - toekomstig* "present - future", *morgen - vandaag* "tomorrow - today" and *huidig - nieuw* "current - new" were also judged as non-opposites but not unanimously but by the majority vote. This suggests that when a non-binary pair evokes a scale but one of the words represents the middle point of the scale, not all participants tend to dismiss such pairs as non-opposites.

Based on manual classification, it was possible to calculate precision scores, see Table 4.7. The Table presents precision scores based on all found pairs with scoring ≥0.9 (col. 2, 3) and based on pairs that co-occurred with each other in the TwNC

Top-k found pairs	Precision scores	Examples of found opposites
50	0.88	*platteland - stad* "country - city", *aanbod - vraag* "supply - demand", *bepaald - onbepaald* "definite - indefinite", *gezond - ziek* "healthy - sick"
100	0.74	*positief - negatief* "positive - negative", *dag - nacht* "day - night", *kansarm - kansrijk* "underprivileged - promising"
150	0.6	*vijand - vriend* "enemy - friend", *vader - zoon* "father - son", *horen - zien* "to hear - to see", *niet-roker - roker* "non-smoker - smoker"
200	0.54	*zout - zoet* "salty - sweet", *win - verlies* "win - lose", *begin - eind* "begin - end", *stijgen - dalen* "increase - decrease"
250	0.5	*vader - moeder* "father - mother", *vrede - oorlog* "peace - war", *letterlijk - figuurlijk* "literally - figuratively"

Table 4.8: Top-k pairs which co-occurred significantly often with scoring $\geq$0.9 extracted with 18 adjective - adjective seeds and examples of found opposites. Precision scores are based on the classification of pairs by three participants.

significantly more often than would be expected by chance (col. 4, 5). The highest precision scores of 0.49 were found for pairs at the highest scoring level of $\geq$0.98. This, and the fact that the largest number of pairs judged as opposites were also among pairs with this score, shows that applied automatic scoring was indicative of antonymy. Although significant co-occurrence improved the precision, especially for pairs at lower score levels, which contained more noise, it is not sufficient to significantly improve the precision.

The highest precision scores achieved for found opposite candidates were lower than those reported in similar studies on hyponym and meronym extraction. Such methods usually use a ranked-based system for evaluation of the results. It might be that our precision scores are lower because they are based on the total number of pairs we find, which in turn can contain non-opposites simply because there is a limited number of opposites overall that can be found. To investigate this, we examined precision scores for the top-k found pairs as is shown in Table 4.8.

When only top-50 best pairs are taken into account, the precision score is as high as 0.88, which is comparable to the state-of the art performance of pattern-based methods for meronym and hyponym identification. Some of the pairs in this list are from the original seed set, but most of them are newly acquired opposites. Also the precision score for the top-100 and top-150 pairs is high, ranging between 0.74 and 0.6. Once the number of pairs has grown to 200, more pairs were judged as non-opposites. Among such pairs were *wedstrijd - training* "competition - training", *jurist - burger* "lawyer - citizen", *bevolking - elite* "population - elite", *Euro - gulden* "Euro - guilder" (current

and previous currency in the Netherlands). These examples have not been discussed in relation to antonymy in any of the theoretical approaches. However, looking at the contexts of such pairs, it is easy to see that they refer to the opposite concepts. For example, *Euro* and *guilder* are used in the contexts of prices comparisons, that is low prices in guilders in the past are compared with high prices in Euro when the Netherlands became a Euro-zone. It is possible to find context to make all these example oppositional. The fact that participants did not recognize such pairs as opposites shows that more context is needed for less typical pairs to be recognized as opposites.

Among unanimously judged non-opposites found in the set of top-50 pairs were *christen - moslim* "Christian - Muslim", *Amerika - Europa* "America - Europe", *bestuur - burger* "government - citizen", which are often contrasted with each other in the newspaper texts; opposites *praktijk - theorie* "practice - theory", *land - stad* "country - city" as well as the pair *lang - middel-lang* "long - middle", which refers to the category of LENGTH but not to the polar opposites of this dimension. It is possible to find contrastive contexts for all of the top-50 pairs, suggesting that the algorithm successfully finds pairs that can be used to find Contrast relations. While manual evaluation shows that indeed some pairs are more typical, conventionalized opposites than others, our results show that in the corpus, canonical well-established opposites behave similar to non-canonical context-dependent opposites and they receive similarly high top automatic scores.

Once all pairs are evaluated by judges, it is possible to assess the completeness of the coverage of opposites in CORNETTO by examining how many of found pairs judged as opposites by two or all three participants were present in this resource and how many of them were linked as opposites. Out of 208 pairs judged as opposites by the majority vote, 73% (152 pairs) had both words listed in CORNETTO. Among those 152 pairs, 35 were linked as opposites (23%). This means that for 77% of opposites listed in CORNETTO, this relationship is not explicitly marked among these pairs. In summary, textual patterns acquired by means of 18 adjective - adjective seeds identified many good opposites, 77% of which are missing in CORNETTO. While all seeds were adjectival, many found pairs were expressed by noun - noun pairs as well as verbs. Thus, identified patterns seem to be general enough to allow for such variation. Besides opposites, many identified pairs were not related by any lexical semantic relation and represented frequently co-occurring pairs like *old - sick*. Among other relations, our method found many correlates (including pairs that expressed identities and locations). The kind of patterns we found will be discussed next.

Dutch	English	Automatic scoring
oost - west	east - west	1
man - vrouw	man - woman**	1
goed - kwaad	good - evil	1
arm - rijk	rich - poor*	1
platteland - stad	countryside - city	1
aanbod - vraag	offer - demand	1
langzaam - snel	slow - fast*	1
noord - zuid	north - south	1
burger - politiek	citizen - political	1
dicht - open	closed - open*	1
kort - lang	short - long*	1
jong - oud	young - old*	1
blank - zwart	white - black	1
hard - zacht	hard - soft*	1
lelijk - mooi	ugly - beautiful*	1
hoog - laag	high - low*	1
Amerika - Europa	America - Europe	1
groot - klein	large - small*	1
goed - slecht	good - bad	1
links - rechts	left - right	1
heet - koud	hot - cold	1
droog - nat	dry - wet*	1
dood - levend	dead - alive	1
bestuur - burger	politics - citizen	1
actief - passief	active - passive*	1
koud - warm	cold - hot*	1
nieuw - oud	new - old*	1
fout - goed	wrong - right*	1
licht - zwaar	light - heavy*	1
wit - zwart	white - black	1
allochtoon - autochtoon	foreigner - indigenous	1
Christen - Moslim	Christian - Muslim	1
bepaald - onbepaald	definite - indefinite	1
praktijk - theorie	practice - theory	1
Moslim - niet-Moslim	Muslim - not-Muslim	1
noordelijk - zuidelijk	northern - southern	1
gezond - ziek	healthy - sick	1
kort - middellang	short - middle long	1
jongen - meisje	boy - girl	1
jongere - oud	adolescent - old	1
kind - ouder	child - parent	0.99
gekozene - kiezer	elected - elector	0.99
lang - middellang	long - middle-long	0.99
niet-werk - werk	not-work - work	0.99
donker - licht	dark - light	0.99
huur - koop	rent - purchase	0.99
dik - dun	thick - thin	0.99
blij - verdrietig	happy - sad*	0.99
hetero - homo	heterosexual - gay	0.99
land - stad	countryside - city	0.99

Table 4.9: Fifty top pairs found with 18 adjective - adjective seeds by means of strictly textual patterns and their automatic scores. A single asterisk indicates that a pair was in the original adjective seed set, a double asterisk indicates that a pair was in a seed set of a different part-of-speech category.

Functional type	Textual patterns	English ecquivalent	Found pairs
Coordinated	niet <ANT> of <ANT> van .	not <ANT> or <ANT> of.	*slecht - goed* "bad - good", *koud-heet* "cold - hot"
	en oud , <ANT> en <ANT> .	and old and <ANT> and <ANT>	*dun - dik* "thin - thick", man - vrouw "man - woman"
Distinguished	de kloof tussen <ANT> en <ANT> in	the gap between <ANT> and <ANT> in	*aanbod - vraag* "supply - demand", *zuid - noord* "south - north"
	de verschil tussen <ANT> en <ANT> ben	the difference between <ANT> and <ANT> is	*Europa - Amerika* "Europe - America", *meisje - jongen* "girl - boy"
Transitional	, van <ANT> tot <ANT> , het	, from <ANT> to <ANT>, the	*kort - lang* "short - long", *links - rechts* "left - right"

Table 4.10: Examples of textual patterns found by means of 18 adjective - adjective seeds; their corresponding types according to Jones [2002] and examples of pairs they extracted.

4.4.1.2 Patterns acquired with 18 adjective - adjective seeds

The novelty of the presented study is that all patterns used to find opposites were identified and scored automatically. Patterns with a score lower than 0.1 and patterns that occurred only once were discarded, other patterns were used for scoring pairs. More than 30k unique patterns were acquired and used to find candidate opposites. Although automatically generated patterns were more specific and diverse than manually identified patterns in Jones [2002], they could be generalized and classified according to different pattern types distinguished in previous work in which they were used to explain and categorize the functions of opposites in discourse. A sample of extracted patterns with the highest scoring and their types according to Jones [2002] is illustrated in Table 4.10.

According to the algorithm, generated patterns could have a minimum length of 3 tokens and a maximum length of 7 tokens. The shortest patterns were four tokens long, the longest - seven tokens. The average length of patterns was six tokens long. Patterns with scoring above 0.5 on average were longer than patterns with lower scores suggesting that more specific patterns found more pairs with higher scores than shorter, more general patterns.

4.4.1.3 Number of seeds and corpus size: adjective - adjective pairs

Above we discussed the results obtained from the entire TwNC of Dutch newspaper texts. In this section, we examine how the size of the corpus and the number of seeds can affect the results. In particular, first, we want to know whether a larger number of seeds improves the recall and precision. Second, given that the algorithm takes a lot of computational power, we investigate how the size of the corpus affects the performance of the algorithm. To address the second point, the same experiments were conducted on two subcorpora: the 100 million words version of the TwNC corpus and the 200

Size	6 seeds		12 seeds		18 seeds	
	Found pairs	Opposites	Found pairs	Opposites	Found pairs	Opposites
100 mln	49	71.4% (35)	71	62% (44)	166	53.6% (89)
200 mln	129	55% (71)	-		-	
300 mln	178	52.3% (93)	-		483	43.3% (209)

Table 4.11: Number of pairs with scoring $\geq$0.9 extracted from data collections of different size (TwNC) by means of adjective - adjective seed sets of different sizes.

million words version of the corpus. The results were compared with the results from the complete TwNC. They are presented in Table 4.11.

Corpus size. Using six seeds, 49 pairs with frequency $\geq$5 and the score $\geq$0.9 were found in the 100 million words subcorpus, 129 pairs were found in the 200 million words subcorpus and 178 pairs were found in the complete TwNC. With 18 seeds, 166 pairs were found in the 100 million subcorpus and 483 pairs in the full version of the corpus. This shows that given the same number of seeds, the size of the corpus positively affects the recall. However, the precision scores are higher in the results for smaller subcorpora. Namely, six seeds led to the precision of 0.75 for the 49 pairs found in the 100 million words subcorpus, 0.57 for the 129 pairs found in the 200 million subcorpus and 0.52 for the 178 pairs found in the full TwNC. Thus, more data leads to higher recall and lower precision. The same effect was found for pairs extracted by means of 18 seeds, where the precision score for the 166 pairs found in the 100 million subcorpus was 0.55 and for the 483 pairs found in the full TwNC - 0.43. Note, that the overall precision for the results found with 18 seeds was lower than for the results with six seeds for both corpora. This brings us to the second question, namely, the role of the size of the seed set.

Number of seeds. Not only larger data collections but also larger seed sets led to higher recall. For example, using the 100 million words subcorpus, six seeds found 49 pairs, 12 seeds found 71 pairs and 18 seeds found 166 pairs. Similar, more seeds found more pairs in the full corpus. The difference in the recall led to lower precision scores for larger seed sets. However, recall that in the previous section we have shown that the precision scores for candidate opposites should be assessed by examining the top-k pairs found by seeds. In other words, rather than examining precision scores based on all found pairs, the number of which differs per seed set and corpus size, it is necessary to take into consideration how well the algorithm performs per each seed set and each corpus version for finding top-k best candidates.

When the number of found pairs is taken into account, the results show that more seeds lead to higher precision. For example, in the 100 million words subcorpus, the precision score for the top-49, that is, all pairs found with six seeds was 0.75, the precision score for the top-50 pairs found with 12 seeds was 0.72 and the precision score for the top-50 pairs found with 18 seeds was 0.8. Interestingly, all 49 pairs found with six seeds were also found with 12 and 18 seeds and all 71 pairs found with 12 seeds were also found with 18 seeds. The ranking of the pairs, however, differed per seed set. For example, ten pairs among 49 pairs found with six seeds were not among top-50 pairs in other two lists. Among such pairs were opposites *populair - traditioneel* "popular - traditional", *modern - oud* "modern - old", *commercieel - publiek* "commercial - public" and context-dependent contrastive pairs *geel - wit* "yellow - white", *bank - verzekeraar* "bank - insurer". Another seven pairs from the top-50 pairs found with 12 seeds were not in the top of the other two lists. These pairs included *breed - lang* "wide - long", *eerste - tweede* "first - second", *dik - dun* "thick - thin", *hogeschool - universiteit* "high-school - university". Fourteen pairs in the top-50 found only by 18 seeds included pairs *burger - politiek* "citizen - politician", *aanbod - vraag* "offer - demand", *daad - droom* "action - dream", *gekozene - kiezer* "elected - elector", *Katholiek - Protestant* "Catholic - Protestant" and others. Another important difference was due to the total number of the original seeds present in the top results. Namely, among 49 pairs found with the set of six seeds, 11 opposites were from the 18 selected canonical pairs, among top-50 pairs found with 12 seeds there were 14 opposites out of 18 and naturally, the top-50 pairs found with 18 seeds had the largest number of original seeds, namely, 16 pairs. Because of this, the top-50 pairs found with 18 seeds had the highest precision score.

Interestingly, given the smallest subcorpus, only 18 seeds found more than 100 candidate pairs. Six seeds needed at least 200 million words subcorpus to achieve similar results. In fact, if we compare precision scores for the top-50 pairs found with six seeds in the all corpora, we can see that again the size of the corpus plays a role in that larger corpora lead to better precision.

On the full corpus, the largest seed set outperformed the set of six seeds for the top-50, top-100 and top-150 pairs. In particular, the precision score of the top-50 pairs found with 18 seeds was 0.878, for the top-100 pairs it was 0.736 and for the top-150 pairs it was 0.605. The set of six seeds achieved the precision of 0.79 for the top-50 pairs, 0.654 for the top-100 and 0.54 for the top-150 pairs. Thus, the largest corpus and the largest seed set together give the best results. Moreover, using 18 seeds with

Scoring	Pairs found with 6 seeds	Pairs found with 18 seeds	Overlap between 6 & 18 seeds
≥0.9	42.2% (603)	41.8% (844)	100% (603)
≥0.8<0.9	19.4% (277)	18.8% (380)	100% (277)
≥0.7<0.8	18.6% (266)	19% (385)	100% (266)
≥0.6<0.7	15% (213)	16% (321)	100% (213)
<0.6	4.8% (69)	4.4% (89)	100% (69)
Total	*1,428*	*2,019*	*1,428*

Table 4.12: Percentage of unique pairs found with six and 18 noun - noun seed sets in a full version of TwNC per scoring level and the percentage of pairs that were found in both sets.

the smallest subcorpus gives worse results than using six seeds on the full TwNC. And more data rather than more seeds gives better precision with the same recall.

4.4.2 Results for noun - noun seed pairs

The seed set with 18 noun - noun pairs extracted twice as many pairs as the set with 18 adjective - adjective seeds. In particular, using a full version of TwNC, a total of 1,428 unique pairs were found with the set of six seeds and 2,019 unique pairs were found with the set of 18 seeds. Pairs found less than five times were dismissed from the results. Pairs with the score below 0.6 were discarded. As can be seen in Table 4.12, fewer pairs were found with a six seed set and, similar to the results for adjective - adjective seeds, all of them were also found with the set of 18 seeds (see the *Overlap* column in the Table). Thus, also for noun pairs, adding more seeds improved the recall of the algorithm. But most of the pairs found with the largest set were also identified by a small number of only six seeds. As a result, the smallest set of seeds identified 71.4% of pairs (with the highest scores of ≥0.9) that were also extracted with set of seeds that was three times larger. Since the results found with the set of 18 seeds contain all pairs found with the set of six seeds, these results will be presented first.

4.4.2.1 Patterns acquired with 18 noun - noun seeds

Out of the total 2,019 unique pairs with the score ≥0.6 that were found with the set of 18 noun - noun seeds, 41.8% (844 pairs) had the score ≥0.9. This is a similar proportion of pairs as compared to the pairs found with adjective - adjective seeds (46% or 483 pairs). The overview of how many of those pairs co-occurred with each

Scoring	Number of pairs	Significant co-occurrence
≥0.9	844	97.9% (826)
≥0.8<0.9	380	96.8% (368)
≥0.7<0.8	385	97.4% (375)
≥0.6<0.7	321	96.6% (310)
Total	*1,930*	97.3% (1,879)

Table 4.13: Number of unique pairs found with 18 noun - noun seeds per scoring level, number of pairs that co-occurred with each other sententially significantly more often than would be expected by chance in the full version of TwNC.

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.9	826	6.6% (40/601)	9.2% (76)	10.2% (84)
≥0.8<0.9	368	2.8% (7/250)	2.7% (10)	3.3% (12)
≥0.7<0.8	375	4.2% (11/259)	3.5% (13)	3.7% (14)
≥0.6<0.7	310	5% (10/197)	4.5% (14)	5.8% (18)
Total	*1,879*	*5.2% (68/1,307)*	3.4% *124*	*6.8% (128)*

Table 4.14: Unique pairs found with 18 noun - noun seeds in TwNC significantly often per scoring level and the number of pairs that were found in one or both of the lexical resources (CORNETTO and *Mijnwoordenboek.nl* (*MWB*)).

other within a sentence in the TwNC significantly more often than would be expected by chance is presented in Table 4.13.

Ninety-seven percent of pairs (1,879) were found significantly more often than would be expected by chance. More pairs at higher score levels co-occurred with each other significantly often. As we will show later, significant co-occurrence improved the precision scores also for pairs found with noun - noun seeds, again demonstrating that significant co-occurrence can be used as an additional means of filtering out non-opposites from the results (see Table 4.16 for details). While necessary, significant co-occurrence was not sufficient as many noun - noun non-opposites found by means of textual patterns also co-occurred significantly often, for example, *glamour - glitter* "glamour - glitter", *e-mailen - surfen* "to e-mail - to surf", *advies - geven* "advice - to give" and others. We discarded pairs without significant co-occurrence from further analysis.

Next, we compared how many of found pairs were opposites according to CORNETTO (Table 4.14). For 69.5% of found pairs with significant co-occurrence (1,307 pairs), both words were present in CORNETTO. Only 5.2% of them, however, were

linked as opposites. More than half (58.8% or 40 opposites) were found among 601 pairs with the score $\geq$0.9. Among opposites with the highest scoring were noun - noun pairs like *dame* - *heer* "lady - mister", *burger* - *militair* "citizen - soldier", adjective - adjective pairs like *blank* - *zwart* "white - black", *groot* - *klein* "big - small", and a verb - verb pair that we used in a verb seed set, namely *verliezen* - *winnen* "lose - win". Seventy-two percent of them were linked as opposites symmetrically. For example, *ondergang* "drawback" was linked as an opposite of *opkomst* "turnout" and *opkomst* "turnout" was linked as an opposite of *ondergang* "drawback". On the other hand, *hoogtepunt* "high-point" was linked as an opposite of *dieptepunt* "low-point" but not the other way around. Among other asymmetric opposites were pairs *lid* - *niet-lid* (member - non-member), *leven* - *dood* "alive - dead", *koper* - *verkoper* "buyer - seller".

As can be seen in Table 4.14, more pairs were identified as opposites by means of the online dictionary *MWB* than CORNETTO. Still, both resources identified fewer opposites among pairs found with 18 noun - noun seeds (128 pairs or 6.8%) than among pairs found with 18 adjective - adjective seeds (157 pairs or 16%), although the noun seed set led to the extraction of twice as many pairs as the adjective seed set (1,879 as opposed to 981 pairs found with 18 seeds). If this evaluation is reliable then these findings are in line with the earlier results of Fellbaum [1995], who suggested that opposites expressed by nouns and verbs are less likely to be found in textual patterns. However, as we will show further, manual evaluation suggests that there are more noun - noun opposites than what is present in the lexical resources.

Opposites identified only in *MWB* contained pairs like *dochter* - *zoon* "daughter - son", *vijand* - *vriend* "enemy - friend", *burger* - *soldaat* "citizen - soldier", *ingang* - *uitgang* "entrance - exit", *lelijk* - *mooi* "ugly - beautiful", *fout* - *goed* "incorrect - correct", *gaan* - *komen* "go - come", and others. Opposites identified only by CORNETTO included pairs *jongen* - *vrouw* "youngster - woman", *klein* - *oud* "little - old", *actief* - *inactief* "active - inactive", *mager* - *vet* "lean - fat". Both resources contained pairs *jong* - *oud* "young - old", *burger* - *militair* "citizen - soldier", *degradatie* - *promotie* "demotion - promotion". Some of these examples reveal more inconsistencies between the range of opposites covered in the resources. In particular, both resources mark such pairs as *man* - *vrouw* "man - woman" and *dame* - *heer* "lady - mister" as opposites. *MWB* also lists the pair *dochter* - *zoon* "daughter - son" as opposites because they are antonymous in relation to gender, but not CORNETTO, although it lists the pair *jongen* - *vrouw* "youngster - woman" as opposites, which is a less typical pair since it is contrastive in relation to gender and age and some theoretical approaches do not recognize

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	67.7% (149)	67.8% (101)	48.5% (294)	79.6% (234)	443
≥0.96<0.98	5.9% (13)	69.2% (9)	15.3% (93)	83.9% (78)	106
≥0.94<0.96	11% (24)	83.3% (20)	13.4% (81)	83.9% (68)	105
≥0.90<0.94	15.4% (34)	79.4% (27)	22.8% (138)	86.3% (119)	172
Total	*26.6% (220)*		*73.4% (606)*		*826*

Table 4.15: Percentage of pairs with scoring ≥0.9 extracted with 18 noun - noun seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

this type of pairs as antonymous. A similar pair would be *daughter - father* where there is a contrast in relation to gender and the kinship relationship. Also the pair *burger - militair* "citizen - soldier" was an opposite according to CORNETTO but the pair *burger - soldaat* "citizen - soldier" was not. These inconsistencies highlight the difficulties of antonym classification, as they show that existing theoretical approaches do not provide clear-cut means to distinguish opposites from non-opposites. The context will be helpful in such cases but none of the traditional theoretical classifications use it to decide whether a pair is antonymous or not.

Taken together, CORNETTO and *MWB* helped to identify 6.8% of found pairs as opposites. To know how reliable this result is, all found pairs were further evaluated by three participants. Since 65.7% of opposites were found among pairs with the score ≥0.9, only these pairs were evaluated by judges (826 pairs in total).

In the evaluation task, participants achieved the same level of agreement as for the results for the adjectival seed set. Namely, they achieved a Fleiss's kappa score of 0.66, which indicates high level of agreement. All results are presented in Table 4.15.

Based on the majority vote of three participants, 26.6% of found pairs (220 pairs), which co-occurred in the TwNC significantly often and had a score ≥0.9, were opposites, and consequently 73.4% of found pairs were non-opposites. Among pairs unanimously judged as opposites were *lelijk - mooi* "ugly - beautiful", *broertje - zusje* "little brother - little sister", *kind - volwassene* "child - grown-up", *man - meisje* "man - girl", *arts - patient* "doctor - patient", *docent - student* "teacher - student", *democraat - republikein* "Democrat - Republican". Among pairs unanimously judged as non-opposites were pairs traditionally regarded as co-hyponyms, for example, *auteur - uitgever* "author - publisher", *haas - koe* "hare - cow", *hond - kat* "dog - cat", *danser - musicus* "dancer - musician", *Duitser - Nederlander* "German - Dutchman", *Christen - Moslim*

"Christian - Muslim". Also the pair *generaal - kolonel* "general - colonel", which is part of a closed set of opposites of military ranks according to Lyons [1977], was unanimously judged as non-opposites by the participants. All these pairs are members of multiple member categories, for example, *dog* and *cat* are members of the category ANIMALS, which also includes words *horse*, *pig*, and others; *dancer* and *musician* can be members of the category ARTISTS together with *writer*, *singer*, and *actor*, and so on. However, the reason why these pairs were extracted and automatically scored as highly likely to be antonymous is due to their frequent co-occurrence *in contrastive patterns*. These patterns provide extra context, in which the multiple member pairs are contrastive. Outside of the context, the opposition or contrast between the pairs is not perceived by the participants. But the fact that we did not extract other word pairs from the same categories (for example, *horse - cat*) suggests that automatically found pairs listed above differ from other members of the same category in their contrastiveness. Similar to more readily recognized opposites such pairs co-occur in contrastive patterns and, therefore, they should be treated as antonymous. Further, this also highlights the context-dependence of opposites.

An important question is then whether unanimously judged non-opposites like *dog* and *cat* are actually non-conventional opposites, as they behave in the corpus similar to the well-established opposites, frequently co-occurring in the patterns of incompatibility like [*between* <*ANT*> *and* <*ANT*>], or whether textual patterns that find these pairs are so strong at indicating contrast that even non-opposites they contain appear to be strongly contrastive.

To answer this question, first, consider the following pairs: *rich - poor*, *fast - slow*, *intercity - stop-train*, *Germany - the Netherlands*, *red - white*. There are two canonical opposites on this list, namely, *rich - poor* and *fast - slow* and three co-hyponyms, namely, *intercity - stop-train* (category TYPES OF TRAINS), *Germany - the Netherlands* (category COUNTRIES) and *red - white* (category COLOURS). Both opposites are readily recognized as such by any theoretical approach to antonymy. One of them was found at the top of the list with the results as it had the maximum automatic score of one (the pair *rich - poor*, as well as, pairs *young - old*, *large - small* and *good - bad*); the other pair though appeared at the end of the list as it achieved the automatic score of 0.61 (the pair *fast - slow*, as well as the pair *hard - soft*). The pair *Germany - the Netherlands* was on the top of the list with the score 0.99, *intercity* and *stop-train* had an automatic scoring of 0.88 and the colour terms *red - white* had a score 0.95.

The pair *rich - poor* was found 495 times, mostly in the pattern type [*between*

<*ANT*> *and* <*ANT*>] but also [*for* <*ANT*> *and* <*ANT*>], [<*ANT*> *as well as* <*ANT*>], [*(of)* <*ANT*> *and* <*ANT*>], [<*ANT*> *or* <*ANT*>], [*from* <*ANT*> *to* <*ANT*>], [*for* <*ANT*> *than for* <*ANT*>] and [*from* <*ANT*> *to* <*ANT*>]. The pair *fast - slow* was found only eight times, twice in the pattern type [*between* <*ANT*> *and* <*ANT*>] and six times in the pattern type [<*ANT*> *and* <*ANT*>]. Thus, although both pairs are canonical opposites, only one of them was found very frequently and in a wide range of patterns. This difference was not due to the patterns' specificity, since most of the textual patterns could contain both adjectives and nouns, for example, in the pattern [*the difference between* <*ANT*> *and* <*ANT*>]. Rather, it seems that these pairs differ in the number of contrastive contexts, in which they appear in the given corpus, in this case, newspaper texts. Because Jones et al. [2007] used Google as their data repository, they were able to identify more than ten contexts for each canonical pair of opposites they investigated. However, our results show that in a specific corpus genre, in this case, newspaper texts, opposites with the same level of canonicity differ as to the number of contexts they share.

The pair *Germany - the Netherlands* was found 49 times, 71% of the time it occurred in the pattern type [*between* <*ANT*> *and* <*ANT*>], as well as pattern types [<*ANT*> *and* <*ANT*>] and [<*ANT*> *as well as* <*ANT*>]. In most of the contexts, the two countries were compared in relation to cultural differences and similarities. The pair *white - red* was found 15 times in different variations of the pattern type [<*ANT*> *and* <*ANT*>] in relation to wine. The pair *intercity - stop-train* was found 11 times, 55% of the time in the pattern type [<*ANT*> *and* <*ANT*>] and the rest of the time in the pattern type [*between* <*ANT*> *and* <*ANT*>]. While all these pairs did not co-occur in a wide range of pattern types, they occurred in reliable patterns that indicate their contrastiveness in the given context. The fact that we find all these pairs in productive patterns suggests that they *are* opposites. There seems to be a continuum with well-established easily recognized opposites that share many contexts across different topics and genres on the one side and pairs that share few contrastive contexts in certain domains on the other side. This continuum is dynamic, allowing pairs to move along both directions.

Pairs that were judged as opposites by the majority vote included *echtgenoot - vrouw* "spouse - wife", *geven - vragen* "to give - to ask", *privé - werk* "private - work", *familie - vriend* "family - friend". Such pairs were accepted by the participants as opposites to a smaller degree than unanimously judged opposites for several reasons. First, most of them were not adjectives and they did not evoke any scales as do gradable

adjective pairs, such as *tall - short*. Second, most of these pairs have more typical, or "better" opposites that are lexically conventionalized. For example, the typical opposite of *wife* is *husband*, the typical opposite of *to give* is *to take*, the typical opposite of *private* is *public*, and the typical opposite of *friend* is *enemy*. Finally, since the pairs were presented to the participants without any context, it might be that some of them were not activated as opposites for some of the participants. This is in line with the findings of Willners and Paradis [2010], who discovered that in an elicitation task in Swedish, when participants are asked to provide the best opposite for a given stimulus word, some words had more than one equally preferable opposites. For example, the word *hot* ("het") elicited the opposite *cold* ("kall") 24 times and the opposite *chilly* ("sval") 20 times. Similarly, the word *coarse* ("grov") elicited the opposite *fine* ("fin") 17 times and the opposite *thin* ("tunn") 14 times. These examples illustrate that for some opposites there is no single best second opposite and the preference will change depending on the context or depending on the sense of the word. Adding context in such cases can help participants to "recognize" non-typical and context-dependent antonymous pairs. Therefore, our results seem to suggest that in the future candidate opposites found in textual pattern should be presented to the participants together with patterns or even sentences in which they were found.

Among pairs judged as non-opposites by the majority vote were pairs *directie - personeel* "management - personnel", *kat - muis* "cat - mouse", *Amsterdam - Rotterdam*, Dutch cities, "Amsterdam - Rotterdam", *Ajax - Feyenoord*, Dutch Football clubs "Ajax - Feyenoord", *arm - been* "arm - leg", *antwoord - probleem* "answer - problem", as well as, pairs like *goud - zilver* "gold - silver", *vandaag - morgen* "today - tomorrow", in which one of the words refers to the middle point of the scale. Interestingly, the pair *past - future*, in which the two words refer to the final points on the scale TIME, was judged by the participants as opposites while the pair *today - tomorrow*, in which the middle point is compared with the end point, was judged as non-opposites. However, *today* and *tomorrow* are often contrasted in newspaper sentences, for example, in relation to the climate change, so that the actions that can be taken today will have consequences for the life tomorrow.

Overall what examples of pairs that did not receive unanimous votes as opposites or non-opposites show is that in many cases the context plays a very important role and participants fail to recognize contrast between pairs when they are presented with bare words. Most canonical and typical opposites are unanimously recognized as such, for example, as pairs *man - woman*, *black - white*. Pairs that have more than one suit-

Scoring level	All found pairs	Precision	Pairs with significant co-oc.	Precision
≥0.98	452	0.29	443	0.3
≥0.96 <0.98	109	0.1	106	0.1
≥0.94 <0.96	106	0.22	105	0.23
≥0.90 <0.94	177	0.18	172	0.18

Table 4.16: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 noun - noun seeds.

able opposite, for example, in relation to the intended sense of the word, and opposites that belong to categories with multiple members are recognized by the participants as opposites by majority vote, for example, *cat - dog*, *private - work-related/public*. Pairs that occupy middle points on evoked scales and/or share few contrastive contexts are judged by the participants as non-opposites by the majority vote and unanimously. These judgements are heavily based on participants' intuitions, as a result, there are inconsistencies in the evaluation. For example, the pair *kind - moeder* "child - mother" was unanimously judged as non-opposites, although these are relational opposites because *X is a mother of Y* implies that *Y is a child of X*. The pair *kind - man* "child - man/grown-up" was judged as non-opposites by the majority vote, although these represent opposites in relation to AGE. The pairs *vader - zoon* "father - son" and *dochter - moeder* "daughter - mother" were judged as opposites by the majority vote.

Although manual classification can be misleading, as participants can miss good opposites, it provides the best available assessment of the results for antonymy so we further relied on manual classification to calculate precision scores. Recall that precision scores are based on the number of pairs that received unanimous votes only. This means that 20.6% of pairs, or 170 pairs that did not receive unanimous votes, were not taken into account. All precision scores are summarized in Table 4.16. To examine the influence of significant co-occurrence in the results, the precision scores are presented not only for pairs with significant co-occurrence (columns 4, 5) but also for all found pairs (columns 2, 3).

Significant co-occurrence improved the precision only slightly and the overall scores are rather low. The highest precision score of 0.3 was found for the 443 pairs with the highest scoring ≥0.98, but this is still lower than precision scores reported for the results with adjectival seed sets: the highest precision score of 0.49 was found for 266

Top-k found pairs	Precision scores	Examples of found opposites
50	0.74	*moeder - vader* "mother - father", *dood - leven* "death - life", *oorlog - vrede* "war - peace", *jong - oud* "young - old"
100	0.59	*vijand - vriend* "enemy - friend", *aanbod - vraag* "supply - demand", *commercieel - publiek* "commercial - public", *oost - west* "east - west"
150	0.5	*verlies - winst* "loss - profit", *platteland - stad* "countryside - city", *huurder - koper* "tenant - purchaser", *neef - nicht* "nephew - niece",
200	0.44	*burger - militair* "citizen - soldier", *gezond - ziek* "healthy - sick" *huur - koop* "rent - purchase", *lelijk - mooi* "ugly - beautiful"
250	0.42	*kwaliteit - prijs* "quality - price", *donker - licht* "dark - light", *broer - zuster* "brother - sister", *privaat - publiek* "private - public"

Table 4.17: Top-k pairs which co-occurred significantly often with scoring $\geq$0.9 extracted with 18 noun - noun seeds and examples of found opposites. Precision scores are based on the classification of pairs by three participants.

pairs with co-occurrence higher than chance. Moreover, these precision scores are lower than those reported in the studies on other relation extraction. One of the possible reasons for this discouraging result can be that the scores are based on too many pairs. Since the most typical opposites were among pairs with the highest scores, and pairs at lower score levels were more likely to be non-typical opposites that did not receive unanimous votes from the judges, we also examined the performance of the algorithm based on the top-k pairs. The results are presented in Table 4.17.

The results in Table 4.17 show that indeed the precision scores are high for the top-50 and top-100 pairs with the precision scores of 0.74 and 0.59 respectively, suggesting that the most typical easily recognized opposites are found among the first hundred found pairs. The number of canonical pairs and those opposites discussed in theoretical linguistics is limited. For example, Jones [2002] obtained a list of 36 canonical adjective, noun, verb and adverb pairs after combining resources from theoretical classifications and psycholinguistic studies. Thus, finding 50 opposites is already a good result, given that there are not so many canonical pairs.

Next to finding canonical opposites, the algorithm extracts non-conventional, context-dependent opposites as well as frequently co-occurring semantically similar words which are not opposites. The precision score for the top-50 and top-100 pairs are comparable to the precision scores reported in the studies on meronym and hyponym extraction although those studies aim at finding hundreds of positive pairs as it is possible to find a hypernym or a hyponym for any noun. Recall that hyponymy is the organizing relation for nouns in WordNet. Not any noun, on the other hand, has an

Dutch	English	Automatic Scoring
man - vrouw	man - woman*	1
begin - eind	beginning - end*	1
moeder - vader	mother - father	1
vakbeweging - werkgever	trade union - employer	1
vakbond - werkgever	trade union - employer	1
kind - vrouw	child - woman/wife	1
werkgever - werknemer	employer - employee*	1
reactie - vraag	reaction - question	1
vader - zoon	father - son	1
arm - rijk	rich - poor**	1
dood - leven	dead - alive**	1
aanval - verdediging	attack - defence*	1
nadeel - voordeel	disadvantage - advantage*	1
dag - nacht	day - night*	1
begin - einde	beginning - end	1
blank - zwart	white - black	1
kind - ouder	child - parent	1
broer - zus	brother - sister	1
hel - hemel	hell - heaven*	1
antwoord - kamervraag	answer - question in the parliament	1
dier - mens	animal - human being	1
feit - fictie	fact - fiction*	1
oorlog - vrede	war - peace*	1
kind - volwassene	child - grown up	1
chaos - orde	chaos - order*	1
jongen - meisje	boy - girl	1
antwoord - vraag	answer - question*	1
justitie - politie	Ministry of Justice - police	1
allochtoon - autochtoon	foreigner - indigenous	1
groot - klein	large - small**	1
kerk - staat	church - state	1
dochter - zoon	daughter - son	1
commentaar - vraag	comment - question	1
dame - heer	lady - mister	1
christen - moslim	Christian - Muslim	1
M*á*xima - Willem-Alexander	M*á*xima - Willem-Alexander	1
Duitsland - Frankrijk	Germany - France	1
jaar - maand	year - month	1
jong - oud	young - old**	1
optimist - pessimist	optimist - pessimist*	1
links - rechts	left - right	1
goed - kwaad	good - evil	1
bedrijfsleven - overheid	business - government	1
bond - werkgever	professional organization - employer	0.99
CDA - VVD	CDA - VVD (political parties)	0.99
vragen - zeggen	ask - say	0.99
hetero - homo	heterosexual - homosexual	0.99
Montenegro - Servi*ë*	Montenegro - Serbia	0.99
leerling - leraar	student - teacher	0.99
oost - west	east - west	0.99

Table 4.18: Fifty top pairs found with 18 noun - noun seeds by means of strictly textual patterns and their automatic scores. A single asterisk indicates that a pair was in the original noun seed set, a double asterisk indicates that a pair was in a seed set of a different part-of-speech category.

opposite. This is also supported by the fact that although seeds expressed by adjectives found fewer pairs, more of them were opposites. Once the number of found pairs reaches 150, the precision goes down to 0.5. Among non-opposites found within the top-200 pairs were not only co-hyponyms and pairs that could be contrastive in specific contexts but also frequently co-occurring noun - noun pairs like *arbeid - zorg* "labour - care", *bericht - commentaar* "message - remark", *antwoord - schriftelijk* "answer - written", and others.

Table 4.18 gives on overview of the top-50 pairs found with 18 noun - noun seeds and their scores. Twelve of the original 18 noun - noun seeds are in the top (for example, *day - night*, *man - woman*). In comparison, 16 of the 18 adjective - adjective seeds were among top-50 pairs found with adjective seeds. Noun - noun seeds also found four pairs from the seed set with adjective - adjective pairs, for example, *dead - alive*, *young - old*, whereas adjective - adjective seeds found only one noun - noun seed pair, namely, *man - woman*. This indicates that canonical opposites expressed by adjectives are among the most frequently occurring opposites found not only with adjective - adjective but also with noun - noun seeds.

When the noun - noun seed pair *man - woman* was identified by adjective - adjective seeds, it was found in 876 patterns. The majority of the patterns were variations of the pattern type [*between <ANT> and <ANT>*]. All variations indicated the contrast between men and women, which was not necessarily perceived by judges outside of the context, that is this pattern type. The differences between men and women in the pattern type [*between <ANT> and <ANT>*] were underlined using numerous variations:

gap ("kloof");
contrast/discrepancy ("tegenstelling");
lifespan ("levensduur");
difference ("verschil");
equilibrium ("evenwicht");
difference ("onderscheid");
division/partition ("verdeling");
solidarity ("solidariteit");
segregation ("segregatie");
power ("macht");
everlasting bond ("eeuwigdurend band");
marriage ("huwelijk").
income differences ("inkomen verschil");
dichotomy ("tweedeling");
breach/gap ("gat");
power ratio ("macht verhouding");
contrast ("contrast");
care ("zorg");
similar/equal treatment ("gelijk behandeling");
inequality ("ongelijkheid");
misunderstanding ("onbegrip");
inequality ("ongelijkwaardigheid");
health ("gezondheid");

Interestingly, most of them have not been mentioned in corpus-based work on opposites and their canonicity, suggesting that automatically generated patterns provide a wider range of contexts for canonical and non-canonical pairs even within a given

pattern type.

The pattern of incompatibility was also frequently found among patterns generated by the pair *man - woman* when it was used as a seed. The second most frequent pattern type for the seed pair *man - woman* was the pattern [<*ANT*> *and/or* <*ANT*>]. While this pattern is usually referred to as generic (Pantel and Pennacchiotti [2006]), since it is very frequent but noisy, the variations of this pattern type found in the corpus by means of canonical seeds were too specific to be too noisy. Often, such patterns were part of a longer construction with other opposites, for example, pattern variations [<*ANT*> *or* <*ANT*> *, young or*], [*Muslim and Christian ,* <*ANT*> *and* <*ANT*>], or [*or old ,* <*ANT*> *or* <*ANT*>]. It would be impossible to come up with such patterns based on researcher's intuition alone. Other pattern types, in particular, [*meer* <*ANT*> *dan* <*ANT*>] "more <ANT> than <ANT>", [*van* <*ANT*> *naar* <*ANT*>] "from <ANT> to <ANT>" and [*zowel* <*ANT*> *als* <*ANT*>] "<ANT> as well as <ANT>" were found infrequently. Among many patterns generated with the pair *man - woman* as a seed were patterns of type [*similarity of* <*ANT*> *and* <*ANT*>], for example, [*de gelijkheid van* <*ANT*> *en* <*ANT*>] "the equality of <ANT> and <ANT>", [*de gelijkwaardigheid van* <*ANT*> *en* <*ANT*> .] "the equivalence of <ANT> and <ANT> .", [*gelijk behandeling van* <*ANT*> *en* <*ANT*> .] "equal treatment of <ANT> and <ANT>", [*gelijk recht van* <*ANT*> *en* <*ANT*>] "equal rights of <ANT> and <ANT>", [*gelijk kans voor* <*ANT*> *en* <*ANT*>] "equal opportunity for <ANT> and <ANT>". But the majority of patterns generated by the seed pair *man - woman* were very specific variations of generic patterns [<*ANT*> *and/or* <*ANT*>], as well as, very specific patterns that cannot be categorized into types, for example, [*,* *eenzaam* <*ANT*> *, trots* <*ANT*>] ", lonely <ANT> , proud <ANT>". What this comparison demonstrates is that (1) when a canonical pair is used as seeds, it finds a wider range of *pattern types*. When the pair is not used as a seed, it is still found, but (2) it is found only in *a large number of variations* of a small number of pattern types, mostly pattern types that indicate differences, contrast or similarities between two words. This observation has direct consequences for studies that use the range of pattern types, in particular, the *breadth of antonym co-occurrence* Jones et al. [2007], as it shows that even canonical opposites do not have to appear in a wide range of pattern types to be identified automatically as good opposites.

Pattern types that indicate differences *and* similarities between two words have a unique capacity to capture the fascinating property of opposites, namely, there simultaneous similarity and difference with one another (Cruse [1986], Willners and Paradis [2010]). The fact that seed sets expressed by adjectives and by nouns both find these

patterns and give them the highest automatic score reflects their strong relation with the nature of antonymy. It is, therefore, interesting to see, which other pairs are found by means of these patterns.

The pair *Willem-Alexander - Máxima* - the names of the prince and his wife in the Dutch Royal family - was unanimously judged by the participants as non-opposites. It was found by the algorithm in 191 pattern variations, mostly in the pattern type [*huwelijk / liefde tussen <ANT> en <ANT>*] "marriage / love between <ANT> and <ANT>", followed by the pattern type [*verbintenis van <ANT> en <ANT>*] "commitment of <ANT> and <ANT>". All participants were native speakers of Dutch and they knew who prince Willem-Alexander and M*á*xima were. Nevertheless, they did not identify any contrast between them, as it is difficult to capture the oppositeness of this pair outside of the context. In contrast, based on the patterns, in which this pair was found, we can conclude that this pair has a similar pattern behaviour as its more general counterpart *man - woman*, but whereas the pair *man - woman* is used universally across different topics and genre, the pair *Willem-Alexander - Máxima* has a local contrastiveness highlighted in the newspaper texts only.

In a similar vein, the pair *Germany - France*, which was found in 90 pattern variations, was used contrastively in the newspaper text, being compared like the pair *man - woman* in relation to the [*equality / consensus / domination in relationship / equilibrium between <ANT> and <ANT>*] ("gelijkheid / consensus / macht verhouding / evenwicht"). Because this contrast is local, that is context-specific, as soon as the pair is presented to the participants outside of the pattern, the comparison and contrastiveness between Germany and France is lost.

The aforementioned examples bring us back to the earlier question, which is, whether the contrastive nature of opposites results in their frequent co-occurrence in the patterns of incompatibility or whether patterns of incompatibility are so contrastive that pairs they contain become incompatible. It seems that the wider is the range of the variations of patterns of incompatibility in which a pair of opposites is found, the more typical (in particular, frequent), conventionalized (lexically) and general (used across different contexts) these opposites are. More local, context-dependent opposites and contrastive pairs co-occur in a smaller range of variations of the pattern types. Then, the range of pattern variations, not pattern types, is indicative of the extent of the canonicity of a given pair.

If antonymy is presented as a continuum, then co-occurrence in a wide range of the variations of the pattern type [*between <ANT> and <ANT>*] indicates canonicity. Pairs

that share fewer contexts and occur in fewer variations are less canonical than pairs that co-occur in a wide range of contexts. Pairs that are very domain-specific are the least canonical, in the sense of the least recognizable as opposites by a naive native speaker. For example, the word *black* is a frequent response in elicitation tasks to the stimulus word *white*, as a result this pair is treated as canonical opposites. The word *red* is an unlikely response to the stimulus *white* as this pair is contrastive in relation to WINE only. This will be reflected in that it will be found in a smaller range of variations of patterns of incompatibility. Finally, the pair *Germany - France* will not be recognized as contrastive outside of the context. Its contrastiveness is context-dependent.

Then, the typicality, or canonicity of opposites should not be measured by the range of pattern types but rather by the range of variations of patterns of incompatibility. And this will vary across corpora genre and style.

The majority of pairs (42 pairs) in the top-50 list were expressed by nouns. Some of them were contrastive only in certain contexts. For example, while the most common opposite of *employer* is *employee*, *trade union* was also among candidate opposites found by the algorithm. While this pairing would not be elicited in a psychological study, corpus evidence shows that *employer - employee* and *employer - trade union* co-occur equally often and in similar contrastive pattern types.

Since the part-of-speech category of found pairs was not taken into account, the best 50 pairs found with 18 noun - noun seeds also contained pairs present among the best 50 pairs found with 18 adjective - adjective seeds. In fact, the overlap between the top-50 pairs found with these two seed sets was 26% (or 30 pairs). Among pairs found in sets expressed by both syntactic categories were pairs *child - parent*, *boy - girl* and others. As the number of pairs increased, also the overlap became larger. As can be seen in Table 4.19, 30% of top-100 pairs found with noun seeds were also found with adjective seeds. Among such pairs were *male - female*, *north - south*, *PVDA - VVD* (names of political parties).

Note that because we did not control for the part-of-speech category of the candidate pairs, resulted in some noise. In particular, the seed pair *vraag - antwoord* “question - answer” (noun - noun) was ambiguous with the base form of the verbs *vragen - antwoorden* “ask - answer” (verb - verb) and the combination *ask - answer* (verb - noun). As a result, this seed identified pattern variations that erroneously extracted the pair *zeggen - vraag* “say - question” (verb - noun), *commentaren - vraag* “comment - question” (verb - noun) and a few other cross-categorical pairs. To eliminate such noise, it is necessary to use textual patterns that contain part-of-speech information

Top-k found pairs	Overlap	Pairs found in both seed sets
50	26% (13 pairs)	rich - poor, child - parent, boy - girl
100	30% (30 pairs)	male - female, PVDA - VVD (political parties), north - south
150	32.6% (49 pairs)	friend - enemy, daughter - mother, public - private
200	33% (66 pairs)	father - son, loss - profit, church - state
250	33.2% (83 pairs)	foreign - Dutch, commercial - public, income - expense

Table 4.19: Overlap of pairs found with 18 adjective - adjective and noun - noun seeds that co-occurred significantly often and had the score $\geq$0.9.

about seeds and candidate pairs. This is the topic of Chapter 5.

Going back to the question as to the completeness of the coverage of opposites in CORNETTO, we can now examine for how many opposites found with noun - noun seeds, both words are present in this resource and how many of them are linked as opposites. Out of 220 pairs judged as opposites by the majority vote, 71.4%, or 157 pairs, had both words in the resource and 10.2% of them (16 pairs) were linked as opposites. Among pairs listed in CORNETTO as opposites 56.8% were expressed by adjectives, for example, *wit - zwart* "white - black", *conservatief - progressief* "conservative - progressive"; 40.5% were expressed by nouns, for example, *dieptepunt - hoogtepunt* "low-point - high-point"; and 2.7% were expressed by verbs, for example, *verliezen - winnen* "lose - win". Among opposites that were missing in CORNETTO were adjective - adjective pairs *betaald - onbetaald* "paid - unpaid", *openbaar - privé* "public - private", *ziek - gezond* "sick - healthy"; noun - noun pairs *huurder - verhuurder* "tenant - landlord", *vriend - vriendin* "(boy)friend - girlfriend", *groep - individu* "group - individual"; and verb - verb pairs, for example, *reageren - vragen* "respond - ask". This illustrates that antonymy is not well represented not only for adjectival pairs but also for other syntactic categories and an automatic method for finding opposites can be directly used to improve this.

4.4.2.2 Patterns acquired with 18 noun - noun seeds

Approximately 46k unique patterns were identified by means of noun - noun seeds and used to find new candidate pairs. While the patterns could be three to seven tokens long, the shortest patterns were four tokens long, and the longest - seven tokens long. The average length of patterns was six tokens long. Similarly to patterns found with adjectival seeds, patterns with scoring above 0.5 on average were longer than patterns

Functional type	Textual patterns	English equivalent	Found pairs
Coordinated	die <ANT> of <ANT> die zowel <ANT> als <ANT> krijg	this <ANT> or <ANT> that <ANT> as well as <ANT> get	*nicht - neef* "niece - nephew" *vriendin - vriend* "girlfriend - (boy)friend"
Distinguished	en zorg/inkomen/(on)gelijkheid tussen <ANT> en <ANT>	the care/income/(in)equality between <ANT> and <ANT>	*man - vrouw* "man - woman", *zuid - noord* "south - north"
Transitional	seizoen van <ANT> tot <ANT> niet van <ANT> tot <ANT> te	season from <ANT> to <ANT> not from <ANT> to <ANT> to	*land - stad* "country - city" *finish - start* "finish - start", teen - top "toe - top"

Table 4.20: Examples of textual patterns found by means of 18 noun - noun seeds; their corresponding types according to Jones [2002] and examples of pairs they extracted.

Size	6 seeds		12 seeds		18 seeds	
	Found pairs	Opposites / Precision	Found pairs	Opposites / Precision	Found pairs	Opposites / Precision
100 mln	172	39.5% (68) / 0.42	179	40.8% (73) / 0.43	238	39.5% (94) / 0.4
200 mln	471	29.9% (141) / 0.28	-	-	-	-
300 mln	603	26.8% (162) / 0.24	-	-	844	26% (220) / 0.23

Table 4.21: Number of pairs with scoring $\geq$0.9 extracted from data collections of different size (TwNC) by means of noun - noun seed sets of different sizes.

with lower scores. Thus, longer, more specific patterns were better, that is more frequent and containing more seed pairs, than shorter, more general patterns.

Among patterns with the highest scoring were patterns [*een verschil van <ANT> en <ANT>*] "a difference of <ANT> and <ANT>", [*voor bij <ANT> dan bij <ANT>*] "over for <ANT> than for <ANT>" and [*meer <ANT> dan <ANT> , zoals veel*] "more <ANT> than <ANT>, such as many". Table 4.20 presents examples of found patterns, their types according to Jones [2002] and pairs they acquired.

4.4.2.3 Number of seeds and corpus size: noun - noun pairs

In this section we examine how the number of noun - noun seeds and the size of the used corpus can affect the results. The summary of the results is presented in Table 4.21.

Corpus size. More data led to extraction of more pairs. For example, six noun - noun seeds found 172 pairs with the smallest subcorpus of 100 million words, 471 pairs with the 200 million words subcorpus and 603 pairs with the full TwNC. Even with the largest set of 18 noun - noun seeds, using the 100 million words subcorpus

led to finding fewer pairs than using the 200 million words subcorpus with the smallest set of six seeds. Namely, six seeds found 471 pairs while 18 seeds found 238 pairs. Thus, more data with fewer seeds give higher recall than more seeds and less data. This implies that given a corpus large enough even a small number of seeds can be used to find many candidate pairs.

As the number of found pairs increases, the precision, on the other hand, goes down from 0.42 (six seeds, 100 million words subcorpus) to 0.24 (six seeds, full TwNC). However, these numbers can be misleading since the overall proportion of found pairs that were classified as opposites by at least two participants differed between corpora of different sizes and more opposites were found in the complete TwNC than in its subparts. It might be that when the same number of top pairs is taken into account, the precision scores for larger corpus sizes are higher than for the smallest subcorpus of 100 million words.

When we compare precision scores for the top-k pairs, we find an interesting pattern. Namely, the results show that the same six noun-noun seeds give the best precision for the top-50 pairs with the smallest subcorpus of 100 million words, give the best precision for the top-100 pairs with the subcorpus of 200 million words and the best precision for the top-150 pairs with the full TwNC. In other words, six seeds and a corpus of no more than 100 million words can be used to find a small number of very reliable typical opposites. As more data is used, the algorithm finds a wider range and a larger number of opposites. As a result, using six seeds, the precision score for the top-50 pairs found in the 100 million words subcorpus (0.75) was higher than the precision score for the top-50 pairs in the 200 million words subcorpus (0.68) and the full TwNC (0.68). The precision score for the top-100 pairs found in the 200 million words corpus (0.53) is higher than the precision scores for the top-100 pairs found in the 100 million words corpus and the full TwNC. And the precision score for the top-150 pairs found in the full TwNC (0.47) was higher than the precision score for the top-150 pairs found in the 200 million words subcorpus (0.45) and the 100 million words corpus (0.44).

Interestingly, the precision score for the top-100 pairs found in the 100 million words subcorpus was higher (0.53) than the precision score for the top-100 pairs found in the full TwNC (0.52). This is because as more data was used, less typical opposites were extracted by the algorithm. Recall that the precision scores are based on pairs that receive unanimous votes, but less typical opposites found by the algorithm with more data are more likely to receive majority vote either as opposites or non-opposites. For example, the pair *daughter - mother* did not receive unanimous vote as an opposite and

the pair *golden - silver* did not receive unanimous vote as a non-opposite. While such pairs are opposites, they are not taken into account when assessing the precision scores because they are not typical and are less likely to receive unanimous votes.

This reflects an intriguing tendency. Namely, as more data is added, in addition to the opposites found with smaller subcorpora, the algorithm finds novel less typical opposites and the algorithm favours these pairs over more typical opposites by ranking non-typical opposites among the best candidate pairs. This again indicates the point we made earlier, namely, that given enough data, the non-typical opposites exhibit a similar corpus behaviour as their typical counterparts. This is an important implication for previous comparative corpus-based studies that examined corpus behaviour of canonical and non-canonical opposites Jones et al. [2007]. In particular, they only compared the differences between pairs like *rich - poor* and *rich - wealthy* but ignored pairs like *trade union - employee* and *mother - daughter*. It would be useful to examine differences between such pairs. They also looked only at a limited number of sentences returned by the Google. Our results show that instead it is critical to keep in mind the kind of corpus used and its size as more data can reveal less intuitive opposites.

Finally, given that the best precision score achieved for the top-150 pairs is still lower than precision scores reported in the studies on automatic extraction of other relations, our findings suggests that the number of potential opposites is constrained by the given corpus in that our method can find the most typical and frequent opposites, for example, *employer - employee* and less frequent opposites that are typical for the given corpus. As a result, the number of potential opposites that can be found by the algorithm is limited. For example, the pair *trade union - employer* is frequently contrasted in newspaper texts but not novels or encyclopaedic texts.

Number of seeds. So far we have mainly discussed the results for the set of six seeds. In relation to the number of seeds used, our results show that more seeds give higher precision scores for corpora of all sizes. In particular, for the top-50 pairs found in the 100 million words subcorpus, 18 seeds led to higher precision (0.79) than 12 and six seeds (0.77 and 0.75 respectively). Similar results were found for the top-100 and top-150 pairs. This indicates that using more seeds is better than using fewer seeds.

In comparison with the results based on adjective - adjective seed sets, we find that the number of seeds for both parts-of-speech has a similar impact in that more seeds give better results, that is higher precision. However, while the results based on adjective - adjective seeds suggested that more data, that is, the full TwNC corpus led to better results for all three sets of analysed top-k pairs than smaller subcorpora, the

results based on noun - noun seeds show that the best 50 pairs can be found in the smallest subcorpus and that larger data repository will lead to better precision when more pairs are considered.

This difference in the results can be due to the nature of opposites expressed by nouns as opposed to opposites expressed by adjectives. Very little is known about the nature of oppositeness of pairs expressed by nouns. Many of noun - noun pairs found by our algorithm have not been previously analysed in the studies based on researcher's intuition, a method that is unlikely to come up with non-typical opposites. Therefore, our results are particularly valuable in this respect as they offer a range of possible opposites that have not been encountered in earlier work on antonymy. Knowing that there is opposition between the pair *trade union - employer* is as useful for NLP applications as knowing that there is opposition between the pair *rich - poor*. This also shows that both types of the pairs should be comparatively studied further by theoretical linguists.

4.4.3 Results for verb - verb seed pairs

Seeds expressed by verb - verb pairs found the least number of candidate pairs. In particular, using a full version of TwNC, a total of 99 unique pairs that co-occurred with each other at least five times were found with the set of six seeds and 216 unique pairs were found with the set of 18 seeds. In comparison, six and 18 adjective - adjective seeds found 503 and 1,049 pairs respectively and six and 18 noun - noun seeds found 1,428 and 2,019 pairs respectively. Thus, the highest recall is achieved with noun - noun seed sets. As is shown in Table 4.22, all pairs found with six seeds were among pairs found with 18 seeds. Pairs with scoring below 0.6 were discarded. The results for the set of 18 seeds will be presented next.

4.4.3.1 Patterns acquired with 18 verb - verb seeds

A total of 196 unique pairs with scoring $\geq$0.6 were found by means of 18 verb - verb seeds. Ninety-six percent of them (189 pairs) co-occurred sententially with each other in the TwNC significantly more often than would be expected by chance (see Table 4.23 for details). Candidate pairs found by means of verb - verb seeds have similar significant co-occurrence rates as pairs found by means of adjective and noun seeds (97.4% and 97.3% respectively). Thus, although verb - verb seeds find fewer pairs,

Scoring	Pairs found with		Overlap between
	6 seeds	18 seeds	6 & 18 seeds
≥0.9	36.4% (36)	34.3% (74)	100% (36)
≥0.8<0.9	16.2% (16)	22.2% (48)	100% (16)
≥0.7<0.8	26.3% (26)	22.2% (48)	100% (26)
≥0.6<0.7	12.1% (12)	12% (26)	100% (12)
<0.6	9% (9)	9.3% (20)	100% (9)
Total	*99*	*216*	*99*

Table 4.22: Total number of unique pairs found with six and 18 verb - verb seeds in a full version of TwNC per scoring level and the number of pairs found in both sets.

they are capable of identifying pairs that co-occur with each other more often than would be expected by chance.

Among pairs with the highest scoring (≥0.9), 96% (71 pairs) co-occurred with each other significantly often. Manual inspection showed that there were no opposites among the three pairs that did not co-occur significantly often. This shows that our assumption that significant co-occurrence can be used as a useful cue to separate non-opposites from the results is true, similar to the results found by means of adjective and noun seeds. As a result, pairs that did not exhibit significant co-occurrence in the full version of TwNC, were removed from the results.

Note also that although all 74 pairs with scoring ≥0.6<0.8 had significant co-occurrence, not all of them were opposites. Some of them were collocations like *to nemen - risico* "take - risk", *to beperken - schade* "reduce - damage". This demonstrates that significant co-occurrence alone is not sufficient for filtering out noise from the results for verb - verb seeds and, therefore, it should be treated as a cue but not as the decisive factor for oppositeness. These results coincide with the findings regarding the pairs found by means of adjective and noun seeds, suggesting that significant co-occurrence is useful to the same extent with pairs found by all three part-of-speech categories.

Next, we evaluated the remaining 189 pairs that co-occurred significantly often using the lexical resources CORNETTO and *Mijnwoordenboek.nl* (*MWB*). According to these resources, only 26% of found pairs, or 49 pairs, were opposites, that is they were marked as opposites in one or both of the resources. The summary of the results is presented in Table 4.24.

In relation to CORNETTO, out of 189 pairs with significant co-occurrence, 89.4%

Scoring	Number of pairs	Significant co-occurrence
≥0.9	74	96% (71)
≥0.8<0.9	48	91.7% (44)
≥0.7<0.8	48	100% (48)
≥0.6<0.7	26	100% (26)
Total	*196*	96.4% (189)

Table 4.23: Number of unique pairs found with 18 seeds expressed by verbs per scoring level, number of pairs that co-occurred with each other sententially significantly more often than would be expected by chance in the full version of TwNC.

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.9	71	32.3% (22/68)	33.8% (24)	37.5% (29)
≥0.8<0.9	44	16.2% (6/37)	16% (7)	17.2% (9)
≥0.7<0.8	48	14.6% (6/41)	8.3% (4)	12.5% (7)
≥0.6<0.7	26	13% (3/23)	15.4% (4)	7% (4)
Total	*189*	*22% (37/169)*	*20.6% (39)*	*26% (49)*

Table 4.24: Number of unique pairs found with 18 verb - verb seeds in TwNC significantly often per scoring level and the number of pairs that were found in one or both of the lexical resources: CORNETTO and *Mijnwoordenboek.nl* (*MWB*).

(169 pairs) had both words listed in this resource but only 22% of them were linked as opposites. While this percentage is overall low, bear in mind that this is the largest proportion of found pairs that are marked as opposites in CORNETTO across three seed sets. In comparison, 13.3% of pairs found with adjective - adjective seeds and 5.2% of pairs found with noun - noun seeds were opposites according to CORNETTO.

Of course, adjective - adjective and noun - noun seeds found many more candidate pairs overall. For example, 5.2% of pairs found with noun - noun seeds stand for 68 opposites marked as such in CORNETTO (out of 1,307 pairs present in CORNETTO out of 1,879 candidate pairs found by the seeds). In a similar vein, 13.3% of pairs found by adjective - adjective seeds stand for 95 opposites marked as such in CORNETTO (out of 712 pairs present in this resource out of 981 candidate pairs found by this seed set). Recall, however, that the best precision was achieved for the top 50 and 100 found pairs. It seems that verb - verb seeds find fewer opposites but they also find fewer pairs overall, which reduces the computing power required by the algorithm and the time by half. For example, for the noun - noun seed set, out of 601 pairs with the score ≥0.9 that were present in CORNETTO, only 40 pairs were opposites according to this

resource. For the verb - verb seed set, out of 169 found pairs present in CORNETTO, 37 pairs were marked as opposites. Given that CORNETTO contains well-established opposites rather than novel pairs, it seems plausible to conclude at this point that verb - verb seeds are better at finding already known opposites than noun - noun seeds.

The largest number of opposites found by means of verb - verb seeds according to CORNETTO were among pairs with the scoring $\geq$0.9 (32.3% or 22 pairs). Twenty of these pairs were symmetric opposites, for example, *blijven* - *weggaan* "to stay - to leave", and two pairs - *afstoten* - *aantrekken* "to reject - to recruit" and "to repulse - to attract" - were linked as opposites asymmetrically. It is not clear why such asymmetry is found in CORNETTO but in order to fix it automatically, it is necessary to identify the sense in which two words are antonymous.

As the scoring lowered, the proportion of opposites among found pairs according to CORNETTO also fell from 16.2% for pairs with scoring $\geq$0.8<0.9 to 13% for pairs with scoring $\geq$0.6<0.7, suggesting that the automatic scoring of pairs reflects the number of opposites among candidate pairs.

Recall, however, that, for the results found with adjective and noun seeds, the Dutch online resource *Mijnwoordenboek.nl* (*MWB*) identified more opposites than CORNETTO. In particular, 91 pairs (19.2%) out of 475 pairs with the score $\geq$0.9 found by means of adjective seeds were opposites according to *MWB* leading to a total of 103 opposites (21.7%) identified by one or by both of the resources. In the results with noun seeds, 76 pairs (9.2%) out of 826 pairs with the score $\geq$0.9 were opposites in *MWB*, leading to a total of 84 opposites (10.2%) identified in one or both of the resources among pairs with the highest score. Thus, for the set of adjective seeds, adding the second resource helped to identify 42 additional opposites to the 61 opposites identified by means of CORNETTO. In a similar vein, for the set of noun seeds, adding the second resource helped to identify 44 additional opposites to the ones marked in CORNETTO. To know if this is also the case for the results found by means of verb - verb seeds, we also examined how many of found pairs were opposites in this resource (see Table 4.24).

In relation to *MWB*, the difference between the number of opposites identified by CORNETTO and by *MWB* was not so large. Namely, out of 71 pairs with the score $\geq$0.9, 24 pairs (33.8%) were opposites according to this resource, resulting in a total of 29 opposites (37.5%) found in one or both of the resources. This means that *MWB* identified seven additional opposites that are missing in CORNETTO. Out of the total 189 pairs found by means of verb seeds, 39 (20.6%) were opposites according to *MWB*, leading to a total of 49 opposites (26%) identified by one or both of the resources. This

means that *MWB* identified 12 additional opposites.

There can be several reasons why fewer pairs were identified as opposites by means of *MWB* and CORNETTO for the results from verb seeds as opposed to adjective and noun seeds. It can be that verb - verb seeds find fewer opposites than seeds expressed by other part-of-speech categories. Using manual evaluation can help to address this question.

It can also be that opposites found by means of verb seeds are missing in the resources as verbal antonymy is not well covered. An initial look at the opposites does not show large differences between opposites present in one of the resources or in both resources. Namely, among opposites found in both resources were verb pairs *loslaten - vasthouden* "to release - to hold", *verbeteren - verslechteren* "to improve - to deteriorate", *duwen - trekken* "to push - to pull". Among pairs present only in *MWB* were verb pairs like *kopen - verkopen* "to buy - to sell", *gaan - komen* "to go - to come"; noun pairs like *verlies - winst* "loss - gain", *gevolg - oorzaak* "result - cause" and adjective pairs like *dood - levend* "dead - alive", *lelijk - mooi* "ugly - beautiful". Among opposites identified as such in CORNETTO but not *MWB* were verb pairs like *doorgaan - stoppen* "to go on - to stop", *landen - opstijgen* "to land - to take off" and noun pairs *antwoord - vraag* "answer - question" and *actie - reactie* "action - reaction". Later in this section, we closely examine the top-50 found pairs (see Table 4.27 for details) to know the overlap and the differences between pairs found in different seed sets. But first, we use manual classification to determine how verb seeds performed overall in comparison to adjective and noun seed sets, by examining how many pairs were opposites according to three participants and by calculating precision scores based on their classification.

Three participants classified all pairs with significant co-occurrence and the score $\geq$0.9 as opposites or non-opposites. They achieved a Fleiss kappa-score of 0.59, which indicates a fair level of agreement. It was nevertheless the lowest inter-annotator score among seeds expressed by different part-of-speech categories. In comparison, participants achieved a Fleiss kappa-score of 0.66 for the evaluation of the results for adjective - adjective seeds and noun - noun seeds. A higher level of disagreement among participants reflects that classification of pairs found by means of verb - verb seeds was more difficult for the participants than classification of pairs found by means of adjective and noun seeds. The results are presented in Table 4.25.

The percentage of opposites by the majority vote in the results for the verb seed set was the highest among three seed sets. Out of 71 pairs, 60.6% (43 pairs) were judged as

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	65.8% (25)	60% (15)	34.2% (13)	53.8% (7)	38
≥0.96<0.98	63.6% (7)	85.7% (6)	36.4% (4)	100% (4)	11
≥0.94<0.96	37.5% (3)	100% (3)	62.5% (5)	80% (4)	8
≥0.90<0.94	57.1% (8)	75% (6)	42.9% (6)	83.3% (5)	14
Total	*60.6% (43)*		*39.4% (28)*		*71*

Table 4.25: Percentage of pairs with scoring ≥0.9 extracted with 18 verb - verb seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

opposites by the majority vote, 70% of which received unanimous votes leading to the precision score of 0.6 which is comparable to the precision scores reported in Pantel and Pennacchiotti [2006]. The other 39.4% (28 pairs) were judged by the majority vote as non-opposites. Seventy-one percent of them received unanimous votes. In comparison, 43.8% of pairs found with the adjective seed set and 26.6% of pairs found with the noun seed set were opposites by the majority vote. However, results from the verb seed set had fewer unanimous votes. In particular, 70.4% of the pairs found with verb seeds, 75.2% of the pairs found with adjective seeds and 79.5% of pairs found with noun seeds received unanimous votes either as opposites or non-opposites, suggesting that there were more pairs that led participants to disagree.

Among unanimously judged opposites, 56.7% were verb - verb pairs, for example, *aantrekken - afstoten* “to attract - to repulse”, *toenemen - afnemen* “to increase - to decrease”, *aankomen - vertrekken* “to arrive - to depart”; 30% were adjective - adjective pairs, for example, *links - rechts* “left - right”, *goed - slecht* “good - bad”, *groot - klein* “large - small” and 13.3% were noun - noun pairs, for example, *oorzaak - gevolg* “cause - result”, *winst - verlies* “victory - loss”. Among unanimously judged non-opposites, 40% were verb - verb pairs, for example, *bieden - loven* “to offer - to praise”, *staan - vallen* “to stand - to fall”, *dreigen - vinden* “to threaten - to discover”; 20% were noun - noun pairs, for example, *kosten - moeite* “costs - inconvenience”, *brons - silver* “bronze - silver”, *kant - wal* “side - shore”; 10% were adjective - adjective pairs, for example, *eerlijk - vrij* “honest - free”, *noodzakelijk - passend* “necessary - relevant” and 30% were verb - noun pairs, for example, *geven - commentaar* “to give - remark” (= “to comment upon”), *trekken - conclusie* “to draw - conclusion”, *stellen - vraag* “to raise - question” (= “to ask”). This shows that cross-categorical pairs found by means

of textual patterns are not good opposites and should be discarded from the results.[1] A simple solution to avoid finding such pairs would be setting up an additional constraint on the results, namely, that candidate pairs must belong to the same part-of-speech category. We examine this approach in Chapter 5.

The most interesting pairs, however, are the ones that did not receive unanimous votes either as opposites or as non-opposites, as they represent the difficult cases, that is pairs that cause disagreement among participants.

Among non-opposites by the majority vote, excluding pairs that did not receive unanimous votes, were pairs *norm - waarde* "norm - value", *kat - muis* "cat - mouse", *vis - vlees* "fish - meat", *deel - part* "part - section" and others. The pair *norm - value* was found in patters that indicate differences and similarities, for example, the pattern [*een kwestie van <ANT> en <ANT>*] "*an issue / problem / question of <ANT> and <ANT>*" and [*maatschappij is <ANT> en <ANT>*] "*society is <ANT> and <ANT>*". These patterns also find many good opposites like pairs *winnen - verliezen* "to win - to lose", *geven - nemen* "to give - to take", *aanbod - vraag* "demand - offer".

The pair *fish - meat* was found in variations of the pattern [*neither <ANT> nor <ANT>*], which is a Dutch expression *vis noch vlees* that means *not knowing what to think or believe*. Because this frequently used Dutch fixed expression contains a very productive pattern, the pair was extracted by the algorithm as a candidate opposite. It might be that participants had difficulties discarding this pair as non-opposites due to its rather contrastive nature, *fish* and *meat* are often contrasted with one another. Also the pair *deel - part* "part - section" is part of the fixed Dutch expression with the same pattern *part noch deel aan iets hebben* "having neither part nor lot in something". Thus, surface textual patterns identified by our algorithm found collocations and fixed expressions, which are not possible to eliminate from the results based on the significant co-occurrence alone.

Other pairs that were judged as non-opposites by the majority vote included pairs that evoked a scale but did not refer to the opposite poles on it, for example, *kleinbedrijf - midden-* "small size company - middle size company", *brons - zilver* "bronze - silver", *goud - zilver* "gold - silver", and so on. Unexpectedly, the seed pair *beantwoorden - vragen* "to answer - to ask" was judged as non-opposites by the majority vote. This example suggests that participants are either less likely to recognize opposites expressed by verbs than by other parts of speech or that these kind of tasks allow for such an error,

[1] Examples of good cross-categorical pairs can be found in Fellbaum [1995] who studied co-occurrence of such pairs as *to begin - endless*.

for example, because participants had to evaluate too many pairs at once, as they were asked to classify 200 pairs per session.

There were also other inconsistencies in participants' judgements. For example, while the pair *horen - zien* "to hear - to see" was judged as non-opposites by the majority vote, pairs *lezen - schrijven* "to read - to write", *drinken - eten* "to drink - to eat", *denken - doen* "to think - to do" were judged as opposites by the majority vote. The differences between such pairs are not clear even if we look at the patterns in which such pairs were found. For example, both *horen - zien* "to hear - to see" and *drinken - eten* "to drink - to eat" have the highest scoring of 0.99. They were found in very similar patterns approximately the same number of times. In particular, *horen - zien* was found 42 times in such patterns as [*meer te <ANT> of te <ANT>*] "more to <ANT> or to <ANT>"; [*te <ANT> en te <ANT> , bijvoorbeeld*] "to <ANT> and to <ANT> , for example"; [*iets * of * ?*] "something to <ANT> or to <ANT> ?". The pair *drinken - eten* "to drink - to eat" was found 40 times in patterns like [*meer te * of te **] "more to <ANT> or to <ANT>"; [*te * en te * , je*] "to <ANT> and to <ANT> , you"; [*iets te * of te * ,*] "something to <ANT> or to <ANT> ,". What this can mean is that the differences between pairs that did not receive unanimous votes either as opposites or as non-opposites are not big and that many of them are non-typical opposites, either because they are not binary or because they do not evoke any scales. But these pairs can indicate contrast and cannot be ruled out as non-opposites. Such pairs can be very useful for many NLP applications, in particular, identification of contrast relationships and should not be discarded as noise (Marcu and Echihabi [2002], Spenader and Stulp [2007]). These examples also point out the importance of a further analysis and classification of such pairs by the theories of antonymy as they show a similar behaviour as well-established opposites. Automatically generated patterns can be of particular interest and usefulness in such cases as they reflect the actual contexts in which such pairs occur in natural language.

The latter point can be clearly illustrated by the following example. Consider the pair *lezen - schrijven* "to read - to write" found by our algorithm. It was found 34 times and obtained a score of 0.99. All patterns, in which it was found, are instances of 'antonym' patterns acknowledged in previous corpus-based studies, for instance, in Jones [2002], and Jones et al. [2007] among others. For example, patterns like [*of hij nu * of * ,*] "or he now <ANT> or <ANT> ,"; [*iedereen kan <ANT> en <ANT> .*] "everyone can <ANT> and <ANT> ."; [*kan <ANT> noch <ANT> .*] "can neither <ANT> nor <ANT> .". While this pair is not regarded as antonymous by theoretical

Scoring level	Total number of found pairs	Precision	Number of pairs with significant co-oc.	Precision
≥0.98	40	0.67	38	0.68
≥0.96 <0.98	11	0.6	11	0.6
≥0.94 <0.96	9	0.37	8	0.43
≥0.90 <0.94	14	0.54	14	0.54
Total	*74*	*0.58*	*71*	*0.6*

Table 4.26: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 verb - verb seeds.

classifications, corpus evidence shows that it occurs in antonym-like contexts. We also found a similar pair *lezen - luisteren* "to read - to listen". However, this pair was found only in one pattern [*mens <ANT> en <ANT>*] "people <ANT> and <ANT>", receiving the lowest score of 0.1 and being discarded as a potential candidate. These examples, taken from real data, demonstrate that non-typical opposites exhibit the same behaviour as canonical opposites, co-occurring in the same types of contrastive textual patterns. Then, it might be that they are perceived as less intuitive opposites due to their overall lower frequency rather than an underlying difference between such pairs and more typical opposites. For example, while the pair *to read - to write* was found only 34 times, the pair *to open - to close* was found 283 times. An equally important factor for establishing "oppositeness" of a pair is the range of pattern types in which it co-occurs. The fact that the pair *to read - to listen* is found only in one pattern reflects that it is a less likely opposite than the pair *to read - to write*.

Of course, there can be contexts in which the pair *to read - to listen* is oppositional. For example, it is common in on-line blogs to have an option of reading someone's interview or listening to the conversation as a podcast. However, this context is not widespread in the newspaper texts, more so, this particular context is not likely to be found in newspaper texts collected between 1999 and 2002. Therefore, for our given corpus this pair is not oppositional.

Based on manual classification, we calculated precision scores, the results are presented in Table 4.26. To show that significant co-occurrence has a positive effect on the results, we present scores for all pairs, as well as only significantly co-occurring pairs. As can be seen, using significant co-occurrence as an additional way of removing non-opposites led to higher precision. The overall precision for pairs with significant

co-occurrence was 0.6 whereas for all pairs - 0.58.

Recall that verbal seeds found the least number of pairs overall. If we take into consideration the number of pairs rather than their scores, the precision score for the top-50 pairs found with verb seeds was higher (namely, 0.66) than that for all 71 found pairs (namely, 0.6). Thus, verb seeds resulted in lower recall and higher precision. However, seeds expressed by adjectives and nouns gave both higher recall and higher precision.

In particular, precision scores for the results from adjective - adjective seeds were 0.88 for the top-50 found pairs, 0.74 for the top-100 found pairs and 0.6 for the top-150 found pairs. Precision scores for the results from noun - noun seeds were 0.74 for the top-50 found pairs, 0.59 for the top-100 found pairs and 0.5 for the top-150 found pairs. These results clearly show that seeds expressed by adjectives give the best precision and recall, followed by nouns. Verb - verb seeds give the lowest recall and precision.

It is interesting to see whether pairs found by means of verb - verb seeds are different from pairs found by means of seeds expressed by adjectives and nouns. The overview of the top-50 found pairs and their scores is given in Table 4.27.

The results show that out of top-50 pairs, 13 were from the original seed set (26%), another four were from adjective - adjective seed set, for example, *cold - hot*, *large - small*), and one pair, namely, *question - answer*, was from the original noun - noun seed set [1]. These results are similar to the results found for adjective and noun seed sets, in which 32% and 24% of top-50 pairs respectively were from the original seed sets.

Recall that as our seeds we used well-established opposites previously studied by theoretical linguists. According to the Co-occurrence Hypothesis (Charles and Miller [1989]) and consequent studies of Jones (for example, Jones [2002]), these opposites are strongly associated with each other and are easily recognized by native speakers of English as opposites because of their high co-occurrence with each other. It is then crucial to understand why is it that not all seed pairs are found among top-50 best candidate opposites? Looking at the remaining pairs, which are not present in the top-50 results, shows that they had lower automatic scores because of their weaker presence in the newspaper corpus in comparison to other seeds. This is an interesting finding as it shows that even a 450 million words corpus of newspaper texts is not sufficient for finding all well-established canonical opposites, previously studied in psycholinguistic

[1] Note, that the noun - noun pair is ambiguous as it also includes all instances of the verb - verb pair *ask - answer*, as their forms coincide.

Dutch	English	Automatic Scores
		1
openen - sluiten	to open - to close*	1
aanvallen - verdedigen	to attack - to defend*	1
beantwoorden - vragen	to answer - to ask*	1
bevestigen - ontkennen	to confirm - to deny*	1
verliezen - winnen	to lose - to win*	1
beginnen - eindigen	to begin - to end*	1
kopen - verkopen	to buy - to sell*	1
verliezen - vinden	to lose - to find*	1
huilen - lachen	to cry - to laugh*	1
dalen - stijgen	to decrease - to increase*	1
dalen - toenemen	to fall - to rise*	1
stellen - vraag	to raise - question	1
vis - vlees	fish - meat	1
exporteren - importeren	to export - to import*	1
aanbod - vraag	offer - demand	0.99
mislukken - slagen	to fail - to succeed*	0.99
dood - leven	dead - alive**	0.99
drinken - eten	to drink - to eat	0.99
horen - zien	to hear - to see	0.99
antwoord - vraag	answer - question**	0.99
oplossen - probleem	to solve - problem	0.99
opstaan - vallen	to rise - to fall	0.99
kant - wal	side - shore	0.99
staan - vallen	to stand - to fall	0.99
lezen - schrijven	to read - to write	0.99
verlies - winst	loss - gain	0.99
doen - zeggen	to do - to say	0.99
verhoren - worden	to interrogate - to become	0.99
blijven - weggaan	to stay - to go away	0.99
dreigen - vinden	to threaten - to find	0.99
huren - kopen	to rent - to buy	0.99
verdedigen - winnen	to resist - to overcome	0.99
kat - muis	cat - mouse	0.99
staan - zitten	to stand - to sit	0.99
hoog - laag	high - low**	0.98
deel - part	part - part / section	0.98
bieden - loof	to offer - praise	0.98
denken - doen	to think - to do	0.98
links - rechts	left - right	0.97
doen - laten	to do - to let	0.97
gaan - komen	to go - to come	0.97
eerlijk - vrij	fair - free	0.97
commentaar - geven	comment - to give	0.97
groot - klein	large - small**	0.97
onderscheiden - vernieuwen	to distinguish - to replace	0.97
koud - warm	cold - hot**	0.96
breken - maken	to break - to make	0.96
conclusie - trekken	conclusion - to draw	0.96
aankomen - vertrekken	to arrive - to depart	0.96
oplossen - vraag	to solve - question	0.95

Table 4.27: Fifty top pairs found with 18 verb - verb seeds by means of strictly textual patterns and their automatic scores. A single asterisk indicates that a pair was in the original verb seed set, a double asterisk indicates that a pair was in a seed set of a different part-of-speech category.

experiments. This means that successful identification of opposites depends not only on the canonicity of pairs, reflected in their significant co-occurrence, but also on the variation and size of the corpus, that is the variety of different genres and topics it covers. While well-established opposites, used in this study as seeds, are strongly associated with each other, as has been shown in psycholinguistic tasks, they are not necessarily equally persistent in the same type of texts. It is important to take this into consideration in the studies on antonym canonicity, as the breadth of co-occurrence might reflect antonym canonicity within a specific topic, or contexts, limited by a given genre.

More than 60% of the top pairs were expressed by verbs, some pairs, however, were ambiguous in that they included instances of the word forms expressed by verbs, as well as, by nouns. For example, pairs *huur - koop* "rent - buy" (verb - verb) and "rent - purchase" (noun - noun); *vraag - antwoord* "question - answer" (verb/noun - verb/noun); *los op - vraag* "solve - question/problem" (verb - verb/noun). While non-categorical pairs, that is pairs, in which words belong to the same part-of-speech categories, do not pose a problem as they represent opposites regardless of their part-of-speech category, it would be useful to eliminate cross-categorical pairs, that is pairs, in which words belong to different part-of-speech categories, as they represent part of fixed expressions and are not opposites. In the next chapter of this dissertation we use a method that takes part-of-speech of the target pairs into consideration, eliminating unwanted cross-categorical pairs.

The pair *kat - muis* "cat - mouse" was in the top-50 candidate opposites. It was found in 20 different variations of the same pattern type [*het spel van <ANT> en <ANT>*] "the game of <ANT> and <ANT>". Our participants classified the pair *cat - mouse* as non-opposites, treating the words in their general sense as co-hyponyms of the hypernym ANIMALS. However, in the pattern "*game of/between <ANT> and <AND>*", *cat* and *mouse* are often used to indicate opposition between two contestants, for example, as in the sentence "*The game of cat and mouse between bloggers and journalists is taking new turns*". One can argue that the expression *game of / between cat and mouse* is an idiom, thus, this pattern is not actually finding opposites. However, this pattern identified such opposites as *to increase - to decrease*, *to give - to take*, *to offer - to ask*, *to ask - to answer*, suggesting that in the newspaper texts, the pattern *game of/between* $Word_1$ *and* $Word_2$ *does* indicate opposition outside of the context of the cat-and-mouse game.

In summary, verb - verb seeds found the least number of pairs. For the top-50

found pairs, they achieved the lowest precision score among the results for seed sets expressed by three different part-of-speech categories. The majority of found pairs in the top-50 results were verbs but the seeds also found nouns, adjectives and cross-categorical pairs. In the next section we will examine the differences between pattern types found by means of verbal seeds as opposed to adjectives and nouns, examining the overlap between pairs found by all three sets of seeds.

Using manual evaluation of found pairs, we can check how many of found pairs, classified by the majority vote as opposites, are present in CORNETTO and linked as such. Out of 43 pairs, 42 pairs (97.7%) had both words listed in this resource but only half of them (52.4% or 22 pairs) were linked as opposites. Sixty percent of these opposites were expressed by verbs and 32% were expressed by adjectives. Out of 20 pairs that were not linked as opposites in CORNETTO, 75% were expressed by verbs, for example, pairs *opstaan - vallen* "to rise up - to fall", *aankomen - vertrekken* "to arrive - to depart", *breken - maken* "to break - to make" and others. This demonstrates that the lexical-semantic relationship of antonymy is still missing among many verbs present in CORNETTO.

In conclusion, our results show that computational lexical resources can strongly benefit from the presented automatic technique for finding opposites by increasing the coverage of opposites among pairs that are already present in the resource by half. This method is particularly beneficial for opposites expressed by verbs as they are not studied as thoroughly as adjectives and nouns.

4.4.3.2 Generated patterns

Interestingly, the results for patterns identified by means of verb - verb seeds differed from the results for patterns found by seeds expressed by adjectives and nouns in two ways. First, there were differences in the total number of automatically discovered patterns between the seed sets, so that verb - verb seeds identified the least number of patterns in comparison to seed sets expressed by adjectives and nouns. Second, there were differences as to the most productive pattern types. This suggests that antonymy expressed by verbs might be less typical than antonymy expressed by adjectives and nouns. Further, verbal antonymy might have a different *main* function in discourse, expressed by a different pattern type.

In regard to the number of identified patterns, 18 verb - verb seed pairs identified fewer than 20k unique patterns. Recall that adjectival seeds identified more than 30k

patterns and noun seeds identified the largest number of patterns, namely, 46k. Thus, in comparison to other seeds sets, verbs identified the smallest number of patterns. The average length of automatically extracted patterns was six tokens long, which is the same as for patterns identified by other seed sets, suggesting that this is the optimal average length for automatically identified patterns, which are specific enough to contain many contrastive pairs expressed by different part-of-speech categories. Similar to the results for other seed sets, shorter patterns received lower scoring, suggesting that longer, more specific patterns found more good pairs than shorter, more general, patterns. This is interesting from the computational point of view as it shows that specificity of patterns, reflected in their average length, is equally important for the results of all three studied part-of-speech categories. And although verb seeds find fewer patterns, the patterns they identify tend to be as specific as patterns identified by adjectives and nouns.

The fact that on average patterns were six tokens long for all seed sets shows that surface textual patterns identified automatically for finding opposites differ from surface patterns identified automatically for finding other lexical relations. In particular, recall that Pantel and Pennacchiotti [2006] report that one of the most productive patterns for finding meronyms is the generic pattern "*X of Y*", which finds a lot of good instances of this relation as well as a lot of noise. Because of that, Pantel and Pennacchiotti used the Web as a repository of additional data to filter out non-meronyms from the results. It seems that this step is not necessary for finding opposites as automatically identified patterns that contain antonymy on average are more specific.

Instead, one of the main causes of the noise in the results on automatically found candidate opposites is erroneous identification of cross-categorical pairs like *to pose - question*. Such pairs, however, can be easily eliminated from the results by controlling for the part-of-speech category of found pairs. This method is explored in detail in Chapter 5.

In regard to the most productive pattern types, in contrast to the results for adjectival and nominal seed sets, where the most productive pattern type was [*difference between <ANT> and <ANT>*], the most productive pattern type for finding opposites identified by means of verb seeds was [*(n)either <ANT> (n)or <ANT>*], as in, for example, *either to sit or to stand*, *neither to eat nor to drink*. Other productive patterns included [*te <ANT> of te <ANT>*] "to <ANT> or to <ANT>" and [*meer te <ANT> dan te <ANT>*] "more to/too <ANT> than to/too <ANT>".[1]

[1]The pattern [*te <ANT> of te <ANT> en*] "too/to <ANT> or too/to <ANT> and" found many verb -

The fact that the pattern of incompatibility [*between <ANT> and <ANT>*] was still the most common pattern type for adjectives and nouns suggests that pattern type variation in our results is not likely to be caused by the specific genre of newspaper texts (since adjectival opposites are still very likely to be found in the pattern [*between <ANT> and <ANT>*]) but rather that opposition expressed by verbs is mostly used in different discourse functions.

Similar to the results with adjective and noun seeds, pairs identified by verb seeds were also found in patterns of incompatibility, as in the sentences *Wij treden slechts op als bemiddelaar tussen koop en verkoop* "We only act as a mediator between the buying and the selling parties". Note, however, that many contrastive pairs in this pattern were ambiguous in that they they could express more than one part-of-speech category in their lemmatized form. For example, in the example above, the algorithm actually found the noun-noun pair *koop - verkoop* "purchase - sales", which coincides with stems in the verb - verb pair *kopen - verkopen* "to buy - to sell".

As a result, some of the automatically identified patterns could contain verbs and adjectives, for example, [*te <ANT> of te <ANT>*] "to/o <ANT> or to/o <ANT>", some contained only nouns, for example, [*het <ANT> of <ANT> van een huis*] "the <ANT> or <ANT> of a house" and some could contain pairs expressed by any of the three part-of-speech categories, for example, [*<ANT> of <ANT>* .] "<ANT> or <ANT> .". The limitation posed by this ambiguity is that pairs, that have the same lemmatized form across different part-of-speech categories, get boosted, based on the joint frequency of their instances expressed by different part-of-speech categories. Such pairs end up in the top results of all three part-of-speech seed sets, limiting the potential for finding new opposites for a given set of seeds of a specific part-of-speech category. This suggests that an approach based of the textual patterns that contain part-of-speech categories of the target can be more suitable for finding a wider range of novel opposites, especially for verbs.

4.4.3.3 Number of seeds and corpus size: verb - verb pairs

In Sections 4.4.1.3 and 4.4.2.3 we have established that the size of the corpus and the number of seeds lead to slightly different results for seed sets expressed by adjectives and by nouns. In particular, for the seeds expressed by adjectives, the results suggest

verb and adjective - adjective pairs as the adverb *too* and part of the infinitive *to* have the same form in Dutch. This is a special case as the same pattern type would not work in other languages, such as, English, where the two have different forms.

Size	**6 seeds**		**12 seeds**		**18 seeds**	
	Found pairs	Opposites	Found pairs	Opposites	Found pairs	Opposites
100 mln	12	83.3% (10)	24	62.5% (15)	27	66.7% (18)
200 mln	27	77.8% (21)	-		-	
300 mln	36	69.4% (25)	-		74	58.1% (43)

Table 4.28: Number of pairs with scoring $\geq$0.9 extracted from data collections of different size (TwNC) by means of verb - verb seed sets of different sizes.

that larger corpora rather than larger seed sets lead to better results, whereas, for the seeds expressed by nouns, the results show that larger seed sets rather than larger corpora lead to better results. In this section we examine how the size of the corpus and the number of seeds affect the results for seed sets expressed by verbs.

The same experiment was conducted on two subparts of the corpus: a 100 million words version of the TwNC corpus and a 200 million words version of the corpus with three sets of six, 12 and 18 seeds. The results are compared with the results from the complete TwNC and the summary is presented in Table 4.28.

Corpus size. Our first result is that more data leads to higher recall. Using six seeds, 12 pairs were found with the 100 million words version of the corpus, 27 pairs were found with the 200 million words version of the corpus and 36 pairs were found with the complete TwNC. Thus, increasing the size of the corpus tripled the number of found pairs, which led to a decrease in the precision. Namely, 83.3% of pairs found in the 100 million words version of the corpus were judged as opposites by the majority vote (precision score of 0.87). Among pairs found with six seeds in the full corpus, 69.4% of pairs were judged as opposites (precision score of 0.71).

Note, that one of the reasons why the precision is lower among pairs found in larger corpora is that more of them are judged as opposites by the majority vote but not unanimously. Thus, our second result is that, similar to the results with noun - noun seeds, a larger corpus leads to the extraction of a wider range of pairs, even with the smallest set of six verb - verb seeds.

Exclusion of found opposites that do not receive unanimous votes from the assessment of the precision and, consequently, from the evaluation of the performance of the algorithm can be misleading. For example, among pairs that were found with six seeds only in the full TwNC were opposites *breken - maken* “to break - to make”, *lezen - schrijven* “to read - to write”, *lenen - sparen* “to borrow - to save”, *gaan - komen* “to go - to come”. Although all these pairs express semantic opposition, they are not

readily recognized as opposites. Nevertheless, all of them can indicate contrast and therefore all of them are useful for many NLP applications that are based on automatically identified opposites. Thus, the fact that these pairs were not unanimously judged as opposites highlights the limitations of our manual evaluation method rather than the performance of the algorithm.

Number of seeds. In relation to the differences in the results related to the number of used seeds, larger seed sets led to extraction of more pairs. In particular, using the 100 million words subcorpus, six seeds found 12 pairs, 12 seeds found 24 pairs and 18 seeds found 27 pairs. At first glance, it seems that fewer seeds led to higher precision, with the precision of 0.87 score for the set of six seeds, the precision of 0.87 for the set of 12 seeds and the precision of 0.7 for the set of 18 seeds. A closer examination of the pairs, however, reveals that the high precision score for the 12 pairs found with six seeds is due to the fact that most of them were from the original seed set, and as the number of seeds increased, the number of novel opposites in the results also increased.

Nevertheless, all verb - verb sets of different sizes gave poor results in comparison to the results found with adjective and noun seeds. For example, even when 18 seeds were used with the complete TwNC corpus, 43 found pairs were judged as opposites by the majority vote as compared to 220 judged opposites in the results with noun - noun seeds and 209 opposites in the results with adjective - adjective seeds. This suggests that seeds expressed by verbs might be less suitable for a pattern-based method for finding opposites than seeds expressed by adjectives and nouns. The possible reasons for this finding are discussed in Section 4.6.

4.5 Results for the corpus of encyclopaedia texts - Wikipedia

In this section we investigate whether and how the genre of the corpus affects the results for a pattern-based algorithm. Often, pattern-based methods are tested on either the newspaper corpora or encyclopaedia texts, or both due to the growing availability of data from these genres. However, they differ from each other and it needs to be established whether that can affect productivity of a pattern-based method. For example, encyclopaedia articles tend to contain repetitive constructions, whereas newspaper texts contain more variations. Repetitiveness can be good, for example, for identification of hyponym - hypernyms by means of patterns like [*X is a kind of Y*] but it might

be that patterns of incompatibility are not frequent enough to find opposites. Studying how our algorithm performs on a corpus of Wikipedia texts will help us to shed light on whether this genre can be used for automatic extraction of opposites by means of patterns.

We present the results of our algorithm on a 127 million words corpus of Wikipedia texts in Dutch. Because this corpus is smaller than the TwNC, only the sets with 18 seeds were used, as it has been shown in Sections 4.4.1.3, 4.4.2.3, and 4.4.3.3 that the larger number of seed pairs can compensate for a smaller corpus size. The overview of all found pairs for the seeds expressed by each part-of-speech category is given in Table 4.29.

Seeds expressed by adjectives gave the best results, followed by nouns. Verb - verb seeds did not find any novel pairs. In comparison to the results for newspaper texts, the algorithm performed poorly with all three seed sets, suggesting that encyclopaedia texts are not suitable for finding a wide range of opposites automatically, as opposed to hyponyms (Hearst [1992]).

The largest number of pairs was found with seeds expressed by adjectives. All 27 pairs with scoring $\geq$0.9 were judged as opposites leading to the precision score of one. They contained all original seeds except for *narrow* - *broad*, as well as the following novel opposites: *noord* - *zuid* "north - south", *oost* - *west* "east - west", *noordwest* - *zuidoost* "north-west - south-east", *man* - *vrouw* "man - woman", *burger* - *politiek* "citizen - politician", *links* - *rechts* "left - right", *negatief* - *positief* "negative - positive", *dierlijk* - *plantaardig* "from animals - vegetable", and *donker* - *licht* "dark - light".

For noun - noun seeds, 15 out of 18 found pairs with scoring $\geq$0.9 were judged as opposites by the majority vote, leading to the precision score of 0.87. However, all of the judged opposites were from the original seed set. Thus, noun - noun pairs failed to find novel pairs.

In a similar vein, verb - verb seeds returned only three pairs. Again all three were from the original seed set.

It seems that the main reason why verb seeds did not find any new pairs is due to their inability to identify any productive patterns, as a result of the infrequency of the seed pairs sententially co-occurring in the set range of pattern length in the corpus.

In contrast to verb - verb seeds, adjective - adjective and noun - noun seeds were able to identify such pattern types as [*between* <*ANT*> *and* <*ANT*>], [*from* <*ANT*> *to* <*ANT*>] and [*either* <*ANT*> *or* <*ANT*>]. All of them were very productive in the

Scoring	Adjective-adjective pairs	Noun-noun pairs	Verb-verb pairs
≥0.9	27 (49%)	18 (53%)	3 (100%)
≥0.8<0.9	8 (14%)	7 (20%)	0
≥0.7<0.8	13 (24%)	6 (18%)	0
≥0.6<0.7	7 (13%)	3 (9%)	0
Total	*55*	*34*	*3*

Table 4.29: Number of unique pairs found in the corpus of Wikipedia texts with 18 seeds expressed by adjectives, nouns and verbs (per scoring level).

corpus of the newspaper texts. However, these patterns were not able to identify good opposites in the encyclopaedia texts besides a handful of well-established pairs.

In summary, while previous studies, such as Hearst [1992], have taken advantage of the repetitive constructions used in encyclopaedia texts, relying on a few productive and very reliable patterns for finding, for example, meronyms, our results show that the lack of variation in Wikipedia is disadvantageous for finding opposites because our method is based on the diversity of more general patterns that were not found in the presented corpus.

4.6 Discussion

The goal of this chapter was to examine whether opposites can be found automatically using automatically generated surface textual patterns like [*difference between <ANT> and <ANT>*] and small sets of seeds, for example, adjectives *rich - poor*. Our main results show that it is possible to automatically identify productive textual patterns using a handful of seeds and that automatically extracted patterns can be successfully applied to finding well-established as well as novel pairs of opposites (see Section 4.6.1). For example, the algorithm found a well-established pair of opposites *white - black*, as well as a less readily recognized pair of opposites *white - red*. In the latter case, the words are semantically opposed only in specific contexts, such as the comparison of wine types.

Further, we found differences in the performance of the algorithm related to the number of seeds we used and the part-of-speech category to which they belonged, as well as the genre and the size of the corpus. In relation to the genre and size of the corpus, our results suggest that the genre of the corpus plays a crucial role in the performance of a pattern-based algorithm. We found that while textual patterns success-

fully found opposites in the corpus of newspaper texts, the same seeds failed at finding opposites in the corpus of encyclopaedia texts (see Section 4.6.3 for further details).

In relation to the number of seeds, we found that sets with six seeds found fewer opposites but most of them were conventional opposites, whereas larger seed sets found more pairs, many of which were non-conventional context-dependent semantically opposed words.

In relation to the part-of-speech category of the seeds, we found that adjective - adjective and noun - noun seeds perform better than verb - verb seeds. This is an important finding in relation to the previous corpus-based studies on antonymy as well as existing corpus-driven body of work that explores pattern-based methods for finding various lexical semantic relations, including meronymy and hyponymy. In particular, previous corpus-based studies on antonymy suggest that opposites expressed by all three part-of-speech categories equally co-occur in the same types of patterns (Jones [2002]). Our findings show that verb - verb opposites are found in patterns less often than adjectives and nouns and that they prefer different types of patterns than opposites expressed by adjectives and nouns (see Section 4.6.2). This indicates that the main discourse function of opposites expressed by verbs is different from the main discourse function of opposites expressed by adjectives and nouns. The direct implication of this result is that a pattern-based method might be less suitable for finding opposites expressed by verbs than it is for finding opposites expressed by adjectives and nouns.

On a larger scale, these results also suggest that it is necessary to explore other methods for finding antonymy expressed by verbs, as opposed to adjectives and nouns. So far, the potential differences in the behaviour of pairs expressed by different part-of-speech categories within patterns have not been thoroughly studied in the existing corpus-based work on other lexical-semantic relations, mostly because the main focus of such studies were relationships that hold exclusively between one particular part-of-speech category, for example, nouns. It is important, however, to explore other methods that might be more suitable for finding relationships between verbs. In Chapter 5, we examine whether controlling for the part-of-speech category improves the results. In Chapter 6 we investigate whether syntactic patterns, which do not take the linear ordering of words into account, perform better at finding opposites expressed by verbs than surface textual patterns.

We will now discuss each of the above-mentioned points in more detail.

4.6.1 Automatic identification of opposites

Effect of the seed sets size. In the corpus of newspaper texts (TwNC, Ordelman [2002]), seeds expressed by all three part-of-speech categories led to the extraction of well-established as well as novel opposites. The best results were achieved with adjective - adjective seeds, followed by noun - noun seeds. Verb - verb seeds found the least number of patterns and pairs.

In particular, the set of 18 seeds expressed by adjectives achieved the precision score of 0.74 for the top-100 pairs and the precision score of 0.6 for the top-150 pairs. The set of 18 seeds expressed by nouns achieved the precision score of 0.59 for the top-100 found pairs and the precision of 0.5 for the top-150 found pairs. The set of 18 seeds expressed by verbs found a total of 71 pairs, achieving the precision score of 0.6.

There can be several reasons why different seed sets performed differently with adjectives outperforming both nouns and verbs and with verbs performing least strongly. Two factors could help adjectives and nouns to outperform verbs. First, adjectives and nouns were often found in patterns, in which they could be substituted with each other but verbs could not. This means that patterns identified by adjectives and nouns got higher automatic scoring and were more reliable than patterns identified by verbs. Second, as has already been mentioned earlier, surface patterns might be less suitable for finding verb - verb opposites than they are for finding adjectives and nouns simply because such patterns are too specific or too short. To know if this is the case, we will examine whether syntactic patterns are more suitable for this task in Chapter 6. But first, we will explore in Chapter 5, whether surface patterns perform better when we control for the part-of-speech category of the candidate pairs, eliminating in this way cross-categorical pairs and eliminating, for example, adjectives from the results with noun - noun seeds.

The main reason why the results from seeds expressed by adjectives achieved higher precision scores than the results from seeds expressed by nouns (although the latter found more pairs) seems be related to the role of antonymy for pairs expressed by different part-of-speech categories. Recall that precision scores are based on unanimous votes only. Adjective - adjective pairs were more likely to be unanimously recognized as opposites, than noun - noun and verb - verb pairs. This was reflected in the Kappa scores, which reflect inter-annotators' agreement. The highest Kappa score among three participants was achieved for the results found by adjective - adjective seeds and the lowest Kappa score was found for pairs found by verbs. Our results are in line with Fellbaum [1995] and Jones et al. [2007] who argue that antonymy is the main orga-

nizing principle of adjectives in the mental lexicon (but not for nouns and verbs), as a result, this relationship was particularly salient for the participants for pairs expressed by adjectives and they were more likely to judge them unanimously as opposites.

Significant co-occurrence. Only pairs, in which both words co-occurred with each other within a sentence significantly more often than would be expected by chance, were considered as candidate pairs.

Significant co-occurrence proved to be a useful additional cue for eliminating non-opposites from the results, especially among pairs with lower scoring, improving precision scores. Yet, it was not a sufficient cue for eliminating all the noise in the results, particularly, for pairs like *advice - to give* (noun - verb), which co-occurred with each other significantly more often than would be expected by chance as part of a fixed expression. This finding is in line with the findings in previous work of Grefenstette [1992].

Note, that in most of these cases, words belonged to different part-of-speech categories. This suggests that cross-categorical opposites discussed in Fellbaum [1995], for example, *end - to begin* (noun - verb), cannot be found by means of textual patterns. In fact, none of the cross-categorical pairs found by the algorithm was antonymous. Nevertheless, knowing that cross-categorical pairs found in textual patterns are not good opposites can be an advantage, as it can be used as an additional cue for filtering out noise, for example, pairs *to pose - question* (verb - noun), *risk - to take* (noun - verb), from the results. In fact, we take this into consideration, treating only pairs of the same part-of-speech category as candidate pairs in the studies described in Chapters 5 and 6.

Among erroneously extracted non-opposites, which co-occurred significantly often in reliable patterns were also pairs like *small - sweet*, *Saturday - Sunday*, and others. While such pairs are not contrastive, they were found in contrastive contexts, showing that our method can be used not only for finding pairs of opposites but also more subtle pairs used in contrastive contexts. This can be particularly useful for automatic identification of contrast discourse relationships. For example, Marcu and Echihabi [2002] argue that automatic identification of contrast can not solely rely on opposites and contrastive discourse markers like *but* and *although*, as many sentences do not contain them. Instead, they suggest that also contrastive pairs are useful for identification of contrast relations. Note, that among such contrastive pairs the authors recommend using cross-categorical pairs. While our pattern-based method presented in this chapter does not find cross-categorical opposites, it can find novel contrastive pairs that have not been previously explored in studies on antonymy.

The advantage of this approach is that it is flexible and finds different contrastive pairs depending on the genre of the corpus. For example, our algorithm found the pair *Dutchman - immigrant*, which is often contrastive in the newspaper texts but not in the encyclopaedia texts. Knowing that *Dutchman* is often contrasted with the *immigrant* in the newspaper texts will facilitate identification of contrast in the same type of texts.

Part-of-speech categories of found opposites. Given that we did not control for the part-of-speech category of found pairs, it is interesting to see whether the algorithm found pairs expressed by all three part-of-speech categories. In fact, some pairs were found by more than one seed set. These pairs were either very frequent canonical opposites, for example, the pair *dead - alive* was found in all three seed sets, or they consisted of words that had the same base form for different part-of-speech categories, for example, the pair *koop - huur* "buy - rent" could be a verb - verb or a noun - noun pair. The largest overlap in the results was found between pairs found with adjective - adjective and noun - noun pairs. As is shown in Table 4.30, 26% of pairs overlapped in the top-50 pairs found by means of adjective and noun seeds. The overlap increased to 30% for the top-100 found pairs. Again, this reflects that adjectives and nouns were likely to co-occur in similar patterns, in which they could often be interchanged for one another.

The overlap in the results between seed sets of verbs and nouns and seed sets of verbs and adjectives was much smaller. Only 8% in top-50 pairs were found by both, verb - verb and noun - noun seeds and 12% in top-50 pairs were found by verb - verb and adjective - adjective seeds. The fact that there was a slightly bigger overlap between verbs and adjectives is due to the productivity of one particular pattern, namely the pattern [*te <ANT> of te <ANT> en*] "too/to <ANT> or too/to <ANT> and" found many verb - verb and adjective - adjective pairs as the adverb *too* and part of the infinitive *to* have the same form in Dutch. This is a special case as the same pattern type would not work in other languages, such as, English, where the two have different forms.

Found differences in pattern preferences are interesting, given that previous studies, particularly Jones [2002] argues that opposites expressed by all three part-of-speech categories do not have different preferences as to the pattern types, because different pattern types in which they co-occur indicate different discourse functions. Then, opposites expressed by all syntactic categories can co-occur in any of the pattern types depending on the textual discourse function they signal. What our results suggest, however, is that opposites expressed by different part-of-speech categories tend to co-occur in pattern types with different discourse functions simply because the patterns found

Top-k found pairs	Overlap verb and adjective seeds	Overlap verb and noun seeds	Overlap adjective and noun seeds
50	12% (6)	8%(4)	26% (13)
100	-	-	30% (30)
150	-	-	32.6% (49)
200	-	-	33% (66)
250	-	-	33.2% (83)

Table 4.30: Overlap of top-k pairs found with 18 seeds expressed by verbs, adjectives and nouns.

with adjectives are often in syntactic positions where a noun is also possible, but a verb is not.

Manual evaluation. Although manual evaluation provided a better way of assessing the results than the usage of existing computational lexical resources, it had its flaws. Namely, when encountering non-canonical opposites, opposites expressed by nouns and verbs, or context-dependent opposites, participants often failed to recognize such pairs as antonymous.

The majority of pairs did not receive unanimous votes. They contained "difficult cases", which could not be clearly categorized using intuition or even existing theoretical classifications, as most of such pairs are not even discussed in the literature. Interestingly, pairs classified by the majority vote as non-opposites appeared to be more similar to opposites than unanimously judged non-opposites.

For example, among such pairs were opposites *Groningen - Maastricht*, names of two cities in the Netherlands, used in a contrastive context to refer to the north and the south of the country as they represent two polar cities on the opposite sides of the Netherlands. While such pairs will not be listed as opposites in any lexical resource, the knowledge that they stand for the north and the south points of the country can be useful for many NLP applications. For example, Mohammad et al. [2008] suggest that opposites are useful in text summarization. They further argue that contrasting words rather than typical opposites are useful for such applications. Also pairs like *Duitsland - Engeland* "Germany - England" (in the context of football), *Kosovo - Servie* "Kosovo - Serbia" can be very useful for such applications.

Lexical resources in evaluation of found pairs. Our results show that opposites expressed by all three part-of-speech categories are under-represented in the most up-to-date available lexical resources in Dutch, including CORNETTO. In particular, 77% of 152 pairs found by means of adjective - adjective seeds and judged by the majority

vote as opposites were present in CORNETTO but not linked as opposite. Even more opposites were missing in CORNETTO from the results identified by means of noun - noun seeds. Namely, 81.4% of 199 opposites found be means of noun seeds had both words present in this lexical resource but not linked as opposites. Although verb - verb seeds found fewer pairs than seeds expressed by adjectives and nouns, the pairs they found are still very useful for enriching lexical resources as 52.4% of 42 opposites found by means of verb seeds and judged as opposites by the majority vote had both words present in CORNETTO but not linked as opposites. This highlights that automatic extraction of opposites can and should be used as a useful way to enrich such computational lexical resources as CORNETTO.[1] These results also demonstrate that CORNETTO on its own cannot be used to reliably evaluate automatically found opposites and manual evaluation is still needed for classification of opposites and non-opposites in the results.

Furthermore, some of the pairs in CORNETTO are linked as opposites asymmetrically, that is, one word in a pair is marked as an opposite of another word but not the other way around. For example, while the word *male* was listed as an opposite of *female*, but *female* was not listed as an opposite of *male*; *to stay* is listed as an opposite of *to leave*, but *to leave* is not the opposite of *to stay*. Such examples seem to be a result of the inconsistencies in the encoding rather than a reflection of an underlying difference between the two opposites. Automatically identified opposites might be used to find such asymmetries in the resource automatically, however, due to the possible multiple senses of each of the words, it might be tricky to automatically add antonymy relationship between asymmetric pairs without a proper automatic word sense disambiguation technique.

Adding the second lexical resource, *Mijnwoordenbook.nl*, did not help either. Although this resource contained slightly more opposites than CORNETTO, the majority were still missing.

Possible number of found pairs. Although the method presented in this chapter outputs an unrestricted number of found pairs, as long as they meet the required criterion (namely, automatic scoring above the set threshold), the results show that the best precision is achieved for the top-200 found pairs. This indicates that the potential number of opposites that can be found might be limited. This is not the case with relations like hyponymy and meronymy, where the number of candidate pairs is potentially

[1] Note, that percentages discussed above do not include those pairs, in which one or both words are not represented in CORNETTO at all. In other words, even more pairs of opposites are missing in CORNETTO.

unrestricted.

4.6.2 *Automatically identified patterns and their types*

We have shown that given a large enough corpus, it is possible to identify useful surface textual patterns automatically using a small set of seeds. Further, our results show that automatically identified patterns are capable of finding otherwise missed opposites.

Automatically identified patterns are more specific than manually-selected patterns. For example, based on the number of productive patterns we found, it seems that many instances of good patterns were missed in the previous corpus-based work.

Automatic identification of productive patterns. In contrast to previous findings, our results also show that opposites expressed by different syntactic categories "prefer", that is primarily co-occur, in different pattern types. Recall that Jones [2002] concluded that opposites expressed by adjectives, nouns and verbs were all likely to co-occur in the same pattern types. Our results show that this is not the case and that opposites expressed by adjectives and nouns occur in different types of patterns than opposites expressed by verbs.

Of course, all opposites are likely to co-occur in different types of textual patterns, especially, in patterns that indicate contrast or incompatibility, such as the pattern [*between <ANT> and <ANT>*]. Most likely for this reason, researchers select these patterns in the majority of studies on antonym canonicity. However, what our findings suggest is that while an approach of taking such patterns as the main criterion for identifying good opposites is sufficient for finding a limited range of opposites, it is not sufficient for identifying a wide range of opposites. This might be the reason why the approach presented in Jones et al. [2007] did not identify *fat*, the most commonly elicited antonym of *thin* in psycholinguistic experiments, although they used the Web instead of a smaller corpus.

4.6.3 *Corpora requirements*

The size of the corpus. The size of the corpus played a role in that more data led to the extraction of more pairs, including opposites. Also larger seed sets gave better results. The smallest seed set used on the largest corpus returned approximately the same number of pairs as the largest seed set on the smallest corpus, resulting in similar precision scores. Thus, when less data is available or the computational power is limited, it is

possible to use more seeds to obtain better results. Note that only half of opposites classified by the majority vote were found by both sets. Another half consisted of different pairs.

Our other finding is that a 450 million words corpus is sufficient for finding not only canonical, well-established opposites like *black - white* but also non-canonical, context-dependent opposites like *red - white*. Findings in previous corpus-based studies on antonymy suggest the opposite, arguing that in order to find non-canonical opposites in patterns, one needs a very large corpus like the Web. In particular, in his pattern-based study, (Jones [2002], pp. 154 - 167) argues that a relatively large newspaper corpus of English, which consisted of approximately 280 million words, was helpful for identifying the most common pairings of opposites used in modern English. For example, Jones found that the most typical opposite of *natural* in the given corpus was *man-made* and not unnatural as is suggested in many dictionaries of English. Based on this, Jones et al. [2007] suggest that corpora of such size (almost 300 million words) can be used only for finding "the relatively conventionalized antonyms" and use the Web in order "... to allow for the development of a more accurate and detailed antonym profile ..." [2007, pp. 136-137]. Our results, on the other hand, show that non-canonical opposites can be found in corpora of 450 million words and that they co-occur in the same type of strictly textual patterns as canonical opposites serving the same discourse functions. Thus, one can extract a range of canonical and non-canonical opposites without relying on the World Wide Web.

The genre of the corpus. The genre of the corpus appeared to play an important role. However, the role of the genre seem to differ as to the relationship. Tjong Kim Sang and Hofmann [2009] reports no significant differences in the perfomance of the algorithm run on newspaper and Wikipedia corpora for finding hyponym-hypernym pairs. What appears to be an advantage for using Wikipedia texts over newspaper texts for meronyms, turned out to be a disadvantage for opposites, especially for noun and verb seed sets. A manual analysis showed that a few patterns were generated and most of them were too general to indicate opposites. Thus, if a target relationship is not indicated by few very reliable patterns, using Wikipedia texts is not optimal due to the lack of variation in constructions used in this genre.

CHAPTER 5

Performance of part-of-speech patterns for finding opposites

In this chapter we present a method for finding opposites based on automatically acquired *part-of-speech patterns* like [*the difference between <ANT/Noun> and <ANT/Noun>*] or [*from <ANT/Adj> to <ANT/Adj>*]. We show that part-of-speech patterns give both high recall and high precision with seed sets expressed by adjectives, nouns *and* verbs. This approach is able to find well-established opposites like *rich - poor*, *buy - sell*, *man - woman* as well as non-typical opposites like "*green - red*". In relation to the performance of surface textual patterns, presented in Chapter 4, the results in this chapter demonstrate that controlling the part-of-speech category of candidate opposites improves the performance of a pattern-based method. This is particularly important for finding pairs expressed by nouns and verbs.

Given that this method does not require any computationally costly preprocessing steps and can easily be applied to vast amount of data, part-of-speech patterns offer a promising solution to automatic extraction of opposites.

5.1 Inspirations for the present study

The primary inspiration for work in this chapter was the study on antonymy extraction presented in Chapter 4, as well as existing work on pattern-based methods for finding such relationships as meronymy and hyponymy (Section 5.1.2). In particular, findings from strictly textual patterns suggest that part-of-speech information of candidate pairs is a useful cue that can be used as an additional constraint to eliminate noise from the results as well as to find less frequent pairs of otherwise missed opposites, particularly for verbs. Therefore, we investigate the role of the part-of-speech information on the performance of a surface pattern-based method for finding opposites, taking into consideration shortcomings of strictly textual patterns (Section 5.1.1).

5.1.1 Limitations of strictly textual patterns in finding opposites

Chapter 4 showed that using only strictly textual patterns, that is surface patterns that do not contain any syntactic information, we can successfully find opposites and contrastive pairs in a large unannotated corpus of newspaper texts. However, such method has two major shortcomings, especially for finding opposites expressed by syntactic categories other than adjectives.

First, all three part-of-speech seed sets tend to find the same pairs at the top of their results. In particular, all sets found the most frequently occurring opposites like *rich - poor*, *man - woman*, and others. This happened because textual patterns generated by seeds expressed by any of the three part-of-speech categories were so general that they tended to contain not only words expressed by the target part-of-speech category but also other categories. For example, the pattern [*between* <*ANT*> *and* <*ANT*>] was found by means of adjective, noun and verb seed sets and in all three cases it extracted the same opposites. Further, because the corpus was lemmatized, it was often not possible to disambiguate between part-of-speech categories of found candidates. For example, a pair *huur - koop* “rent - buy” could be an instance of a noun - noun pair or a verb - verb pair. As a result, less frequent pairs of a given part-of-speech category got lower scores and were dismissed while the same, most frequent pairs prevailed in the results of each part-of-speech seed set.

The second shortcoming of using strictly textual patterns is that they also found cross-categorical pairs, such as *moeite - kosten* “inconvenience - to cost” (noun - verb). Words in these pairs were usually part of fixed expressions, contributing noise.

In this chapter, we present a method that can deal with both shortcomings. In particular, instead of strictly textual patterns, we introduce surface patterns that contain part-of-speech categories for target word pairs. We will refer to such patterns as *PoS patterns*. For example, instead of a strictly textual pattern [*between <ANT> and <ANT>*], we will generate a surface PoS pattern [*between <ANT/ADJ> and <ANT/ADJ>*] with adjective - adjective seeds, a surface PoS pattern [*between <ANT/NOUN> and <ANT/NOUN>*] with noun - noun seeds and a surface PoS pattern [*between <ANT/VERB> and <ANT/VERB>*] with verb - verb seeds. In this way, a PoS pattern [*between <ANT/ADJ> and <ANT/ADJ>*] will only find candidate pairs expressed by adjectives like *rich - poor*, while a PoS pattern [*between <ANT/NOUN> and <ANT/NOUN>*] will only find candidate pairs expressed by nouns like *man - woman*. The main advantage of surface PoS patterns is that, while they can deal with the aforementioned limitations of strictly textual patterns, PoS patterns require only minimum syntactic preprocessing (shallow parsing, particularly, part-of-speech tagging). This can be executed at the considerably lower processing costs in a much shorter period of time. In comparison to the dependency patterns, which are becoming increasingly popular in relation extraction, shallow parsing can be applied to a vast amount of data. For example, Tjong Kim Sang and Hofmann [2009], who compared the performance of PoS patterns with dependency patterns aimed at finding hyponym-hypernym pairs, report that they had to refrain from using the most recent, available corpus of Dutch Wikipedia texts in their study because it would take 296 days to perform the full syntactic parsing on the corpus on a single processor machine. In comparison, it would take one hour to tag the same data with part-of-speech information needed for generating surface PoS patterns. In addition, unlike syntactic parsing, PoS tagging is extremely accurate.

Surface patterns with part-of-speech information have been widely used in relation extraction, as it is often the case that certain lexical semantic relations are expressed by a particular part-of-speech category. For example, studies that deal with automatic extraction of hypernym-hyponym pairs are interested only in finding pairs expressed by nouns. Still, there are differences between PoS patterns used in different studies on automatic relation extraction. These differences and their impact on the results will be discussed next.

5.1.2 *Previous studies on surface part-of-speech patterns*

As has been already discussed in Chapter 4, a pattern-based method for relation extraction was originally proposed by Hearst [1992], whose main goal was to find hyponyms like *tulip - flower*, *broken bone - injury*, *chair - furniture*. Hearst proposed to use manually crafted surface patterns like [*<Word1> is a kind of <Word2>*] to find candidate pairs. However, since hyponymy is the 'type of' relation that predominantly holds between noun - noun pairs, Hearst focused at finding nominal pairs only.[1] To do that, surface patterns were modified to include information about the part-of-speech category of candidate pairs. Therefore, a modified surface pattern [*<Word1/Noun> is a kind of <Word2/Noun>*] would only find candidate pairs *Word1 - Word2* expressed by nouns. One of the main difficulties such patterns were facing is the fact that nouns are often modified by determiners, quantifiers, adjectives and so on. As a solution, Hearst only extracted pairs in which both nouns were not modified or they were modified by a small set of listed determiners.

A similar approach was used by Tjong Kim Sang and Hofmann in their studies on finding hyponyms in Dutch. In the first study, Tjong Kim Sang and Hofmann [2007] used automatically acquired surface patterns like [*such <Word/N-pl> as <Word/N-sg>*] to find candidate hyponyms. Using a 300 million words version of the Twente Nieuws Corpus of Dutch newspaper texts, all possible patterns for each word pair in each sentence in the corpus were automatically identified. The maximum length of a pattern was set to four. Following Snow et al. [2005], Bayesian Logistic Regression (BLR) was used to determine patterns indicative of hyponymy. All found noun - noun pairs that were also present in the Dutch WordNet and were associated with at least five hyponym-hypernym patterns, were stored in a dataset as positive or negative evidence, depending on whether a given pair was linked by hyponymy relation in Dutch WN. A classifier was trained using BLR and the performance was tested by 10-fold cross-validation. They achieved a precision score of 0.36.

In their next study, Tjong Kim Sang and Hofmann [2009] also automatically generated textual patterns with PoS information using the TwNC but this time patterns could contain noun phrases, for example, [*such <Word1/NP> as <Word2/NP>*] where the NP could contain a determiner / adjective / noun / proper name. The final token of the matched noun phrase was treated as a candidate noun (the head). Again for each noun

[1] Lyons (1977) notes that for some syntactic categories hypernyms are often of a different syntactic category than the hyponyms. For example, adjectives *happy/sad* share a nominal hypernym *emotion*. He refers to such pairs as *quasi-hyponyms*.

pair in each sentence in the corpus, the algorithm automatically generated surface PoS patterns with the maximum length of five tokens. Next, for each pattern that contained at least five different noun pairs, information was stored as to how many of pairs found in the pattern were hyponyms according to the Dutch WordNet and how many were not. Only noun pairs that were found by at least five different patterns were considered. Finally, a machine learning system was trained (using BLR) to predict whether two given nouns were hyponym-hypernym pairs based on the patterns in which they co-occurred. Evaluation was performed by 10-fold cross-validation. Using the TwNC the precision of 43.1% was achieved.

Both studies on hyponymy extraction in Dutch suggest that a method with PoS patterns leads to a precision between 36% and 43%. These scores are higher than a 20.7% precision reported for the results from a similar approach examined by the authors in the same paper that was based on dependency patterns. A manual analysis of errors (pairs that were missed by PoS patterns but found by means of dependency patterns (81 pairs) and vice versa (104 pairs)) revealed that 64% of good pairs that were not found by means of dependency patterns were due to parsing errors. In return only 12% of pairs missed by PoS patterns were due to PoS tagging errors. This is because while the state-of-the-art available parser for Dutch, Alpino, has a labeled dependency accuracy of 89% (van Noord [2006]), the part-of-speech tagger used in the study achieves an accuracy of 96%. Given that shallow PoS parsing is an inexpensive preprocessing step in terms of both processing costs and time, the authors argue that a PoS pattern-based method gives the optimal performance.

Interestingly, Tjong Kim Sang and Hofmann [2009] report that 48% of hyponym-hypernym pairs missed by the PoS patterns were not found in their study because full parsing was required. For example, PoS patterns could not identify hypernym-hyponym pairs *illness - scurvy* and *illness - beriberi* in the construction "*...illnesses caused by vitamin deficits, like scurvy and beriberi*" because these words occurred too far from each other, given that the maximum length of PoS patterns was set to five. Such examples with long-distance dependencies illustrate the potential limitation of PoS patterns for finding opposites. In such cases, full parsing and dependency patterns provide a better alternative. This topic is fully addressed in Chapter 6.

Being a central relation in many theories of lexical organization, including the approach taken in the WordNet project, many more pairs are linked by the hyponymy relation than by antonymy, so the recall for finding opposites will be much lower. However, we expect that in comparison to strictly textual patterns, surface PoS patterns will

lead to higher precision in antonym extraction. Recall that the precision scores for top-100 pairs found by means of textual patterns were 0.736 for 18 adjective - adjective seeds, 0.586 for 18 noun - noun seeds and 0.517 for 18 verb - verb seeds (only 80 pairs in total were found with verb - verb seeds). Thus, we expect a precision between 0.5 and 0.8 for results found by means of PoS patterns. Note, that while evaluation of the results in Tjong Kim Sang and Hofmann [2009] was based on the computational lexical resource alone, for the evaluation of our results we use not only lexical resources but also manual evaluation. We show that evaluation of the results for antonymy based on lexical resources alone is not accurate because many opposites have both words listed in, for example, CORNETTO, they are not linked by the antonymy relation. We also show that manual evaluation of candidate pairs of opposites has its drawbacks as well and that the best precision is achieved when both means of evaluation are combined together.

5.2 Assumptions

Based on the results in existing studies discussed above, we have the following assumptions for the results of the algorithm that uses part-of-speech patterns for finding opposites:

1. **Automatic identification of opposites**:
    - opposites found automatically will be expressed by all three part-of-speech categories;
    - well-established canonical opposites will be found in a wider range of automatically identified pattern types than non-canonical opposites;
    - cross-categorical pairs will not be found by the algorithm;
    - in comparison to strictly textual patterns, PoS patterns will lead to lower recall (fewer pairs) but higher precision (less noise).
2. **Automatic identification of PoS patterns**:
    - given a large enough corpus, it is possible to identify useful surface part-of-speech patterns automatically;
    - automatically generated part-of-speech patterns can successfully find good opposites;

- in comparison to strictly textual patterns, PoS patterns will find a wider range of opposites for each part-of-speech category. This method will be particularly beneficial for opposites expressed by nouns and verbs, for which strictly textual patterns were not very productive.

5.3 Method

5.3.1 Corpus

We used the Twente Nieuws Corpus (TwNC, Ordelman 2002) which was also used in Chapter 4 and in the studies of Tjong Kim Sang and Hofmann [2009], Tjong Kim Sang and Hofmann [2007]. This corpus is made up of approximately 450 million words taken from newspaper texts. The corpus we used was preprocessed by means of the Alpino parser (van Noord [2006]). The corpus was tokenized (punctuation marks were separated from words and sentence boundaries were identified), tagged with part-of-speech categories and lemmatized (all words were reduced to their base forms).

5.3.2 Seeds

The sets of seeds used in this study were the same as the ones described in Chapter 4 but now they all contained tags with part-of-speech information, for example, *mooi<adj> - lelijk<adj>* "beautiful<Adjective> - ugly<Adjective>". Three sets with six, 12 and 18 seeds were compiled for each of the three part-of-speech categories: adjectives, nouns and verbs. A complete list of seed pairs was summarized in Table 4.1.

5.3.3 Algorithm

The algorithm is very similar to the one used for finding opposites by means of strictly textual patterns. First, the corpus was digitized - converted into numbers - to improve the efficiency of the algorithm and all possible PoS tags were collected. Next, all sentences that contained both halves of any of the predefined seed pairs of the specific part-of-speech category were extracted from the corpus. Based on these sentences, all possible surface patterns of consecutive words were generated, with a minimum pattern length of three and a maximum length of seven tokens. Thus, for an adjective - adjective seed pair *rich<adj> - poor<adj>*, only sentences in which these words were

tagged as adjectives were used to generate patterns, discarding sentences where *rich - poor* were tagged as nouns.

Once all surface PoS patterns were identified, the corpus was searched through again for all occurrences of the patterns with the wildcard tokens ("<-1>") given that words they contained had the same part-of-speech tag as the seed pairs. Patterns that were found only once were eliminated. The rest of the patterns were automatically scored:

$$S\_pattern_i = \sin(\frac{F_i \times \frac{\pi}{2}}{N_i + c}) \tag{5.1}$$

where c was a small constant to prevent the denominator of the above formula to be zero. After preliminary testing, the value of c was set to 5. Patterns with a score lower than a set threshold τ set to 0.1 were dismissed. Finally, based on the scoring of patterns, found pairs found in the wildcard positions were also automatically scored and sorted in the ranked order:

$$AntS(pair_j) = 1 - \prod_j (1 - S(P_i))^{C_{ij}} \tag{5.2}$$

where S(P_i) is the score of pattern$_i$ and C_{ij} is how often the j-th pair occurred in the i-th pattern.

Extracted pairs that consisted of numerals, punctuation marks or frequent words from the stop list were discarded. Also pairs with a score lower than 0.6 and pairs that were found fewer than five times were dismissed. The rest of the pairs are discussed in the next section.

5.4 Results

In this section we present all results obtained from the Twente Nieuws Corpus of Dutch (Ordelman [2002]) in detail. The results are presented separately, first for seeds expressed by adjectives (Section 5.4.1), then nouns (Section 5.4.2) and finally verbs (Section 5.4.3). Our findings demonstrate that a method based on surface patterns with part-of-speech information gives high precision and high recall, outperforming a similar method based on strictly textual patterns that does not use part-of-speech information (Chapter 4). The strength of PoS patterns is especially clear for opposites expressed by nouns and verbs.

Scoring	Pairs found with		
	6 seeds	12 seeds	18 seeds
≥0.9	46.8% (297)	51.1% (1,326)	50.1% (1,641)
≥0.8<0.9	19.2% (122)	21.4% (555)	18.7% (613)
≥0.7<0.8	16.2% (103)	14.9% (387)	16.6% (543)
≥0.6<0.7	13.4% (85)	10.4% (271)	11.3% (370)
<0.6	4.4% (28)	2.2% (56)	3.3% (108)
Total	*635*	*2,595*	*3,275*

Table 5.1: Total number of pairs found with six, 12 and 18 adjective - adjective seeds by means of PoS patterns in TwNC presented per scoring level.

5.4.1 Results for adjective - adjective seed pairs

Using a full version of TwNC, 635 unique pairs with frequency ≥5 were found with a set of six seeds, 2,595 pairs were found with a set of 12 seeds and 3,275 pairs with a set of 18 seeds. Recall, that with the largest set of 18 adjective - adjective seeds, textual patterns found a total of 1,049 pairs (see Table 4.3). Thus, our first finding is that, contrary to our expectations (see Section 5.2), PoS patterns found many more pairs than strictly textual patterns. The summary of how many pairs were found with each seed set per scoring level is presented in Table 5.1.

More seeds led to finding more pairs and all pairs found with smaller seed sets were also found by the largest set of 18 seeds. Thus, a larger seed set found more pairs including all pairs found by smaller sets. For this reason, we will now discuss the results found with 18 adjective - adjective seeds in more detail than results from other seed sets, although we will also compare the performance across all seed sets later on in this section. This comparison is of particular interest because while six seeds found 297 pairs with scoring ≥0.9, 12 and 18 seeds found at least four times more pairs at the same scoring level. This means that more seeds lead to higher recall, but it is important to answer how this affects the precision of the algorithm.

5.4.1.1 Found pairs

As is shown in Table 5.2, out of the total 3,167 found pairs with scoring ≥0.6, 92.5% (that is 2,927 pairs) co-occurred sententially with each other in the newspaper corpus significantly more often than would be expected by chance. More pairs had significant co-occurrence at higher scoring levels, reaching 95.4% for pairs with score ≥0.9. Among pairs that did not co-occur significantly often were *exotisch - modern* “exotic -

Scoring	Number of pairs	Significant co-occurrence
≥0.9	1,641	95.4% (1,563)
≥0.8<0.9	613	88% (538)
≥0.7<0.8	543	91% (493)
≥0.6<0.7	370	90% (333)
Total	*3,167*	*92.5% (2,927)*

Table 5.2: Total number of pairs found by means of PoS patterns with 18 adjective - adjective seeds(column 2), and the percentage of pairs that co-occurred with each other within a sentence significantly more often than would be expected by chance in the TwNC.

modern", *donker - goed* "dark - good" and other similar pairs, showing that sentential significant co-occurrence of found pairs is a useful simple technique for filtering out noise from the results.

However, manual inspection showed that among found pairs that co-occurred significantly often there were also non-contrastive pairs like *hip - jong* "hip - young", *brutaal - klein* "cheeky - small", *beroemd - rijk* "famous - rich". Most of such pairs were found in the results with scoring <0.9. This means that significant co-occurrence is not sufficient for identification of opposites. As a result, we discarded pairs with scoring <0.9.

Out of 1,563 pairs with significant co-occurrence and a score ≥0.9, 73.9% had both words present in CORNETTO (1,155 pairs) but only 13.8% of them (that is 160 pairs) were marked as opposites (see Table 5.3 for details). Canonical opposites like *dom - slim* "stupid - clever", *extern - intern* "external - internal", *negatief - positief* "negative - positive" and so on were among 18.8% of opposites with the highest score above 0.98 (123 pairs). This shows that automatic scoring reflects antonymicity of pairs in the results, so that most readily recognized pairs are on the top of the results. Other opposites according to CORNETTO included pairs *oud - vers* "old - fresh", *mooi - slecht* "nice - bad", *druk - rustig* "busy - calm", *schoon - vies* "clean - dirty".

Almost 22% of opposites in CORNETTO were asymmetric opposites, that is only one of the words in a pair was linked as an opposite with the other and not the other way around. For example, a morphologically-related pair *belangrijk - onbelangrijk* "important - unimportant" was linked symmetrically whereas a similar pair *gevoelig - ongevoelig* "sensitive - insensitive" was not. This illustrates that the asymmetry is not intentional and should be systematically corrected in the lexical resource.

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.98	877	18.8% (123/655)	13.8% (121)	19.6% (172)
≥0.96<0.98	207	9.9% (16/161)	3.9% (8)	9.2% (19)
≥0.94<0.96	169	6.3% (8/127)	5.3% (9)	7.7% (13)
≥0.90<0.94	310	6.1% (13/212)	3.9% (12)	6.4% (20)
Total	*1,563*	13.8% (160/1,155)	9.6% (150)	14.3% (224)

Table 5.3: Pairs with scoring ≥0.9 found with 18 adjective - adjective seeds in TwNC by means of PoS patterns and the number of pairs that were found in one or both of the lexical resources (CORNETTO and *Mijnwoordenboek.nl* (*MWB*)).

MijnWoordenboek.nl (*MWB*) identified fewer opposites (150 pairs in total) and it was not possible to estimate how many of found pairs were overall present in this online dictionary. Interestingly, the overlap between identified opposites in the two resources was not big. Namely, out of the total 224 identified opposites (14.3% of found pairs), 38.4% (86 pairs) were opposites in both resources. Among such pairs were opposites *nieuw* - *oud* "new - old", *jong* - *oud* "young - old", *antiek* - *modern* "antic - modern". Another 28.6% (64 pairs) were opposites only in *MWB*. For example, pairs *hedendaags* - *ouderwets* "contemporary - outdated", *modern* - *ouderwets* "modern - outdated", *enorm* - *klein* "enormous - small". And 33% (74 pairs) were opposites only according to CORNETTO. They included pairs *oud* - *recent* "old - recent", *modern* - *oud* "modern - old", *klein* - *oud* "little (young) - old", *oud* - *vers* "old - fresh". These examples show that each resource lacks certain combinations for the same candidate pairs, for example, *old* in *MWB* is mostly contrasted in its sense of time whereas in CORNETTO, it is contrasted also in relation to the age and freshness. This comparative assessment points out inconsistencies between two resources as well as the benefits of using automatically found pairs for a consistent improvement of the coverage of opposites in contemporary lexicons.

In relation to the results found by means of 18 adjective seeds with strictly textual patterns (see section 4.4.1), recall that 157 pairs found with textual patterns and 224 found with PoS patterns were identified as opposites in one or both of the resources. Thus, there was a 43% increase in the number of opposites in the results found by means of surface PoS patterns. Given that with the adjective seed set, PoS patterns found adjective - adjective pairs only, whereas strictly textual patterns found also noun - noun pairs, we can preliminary conclude that surface patterns with part-of-speech information outperform textual patterns in finding opposites expressed by adjectives.

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	65.8% (340)	75.6% (257)	51.3% (537)	85.5% (459)	877
≥0.96<0.98	11.2% (58)	77.6% (45)	14.3% (149)	89.9% (134)	207
≥0.94<0.96	7.9% (41)	73.2% (30)	12.2% (128)	93% (119)	169
≥0.90<0.94	15% (78)	61.5% (48)	22.2% (232)	89.2% (207)	310
Total	*517 (33%)*		*1,046 (67%)*		*1,563*

Table 5.4: Percentage of pairs with scoring ≥0.9 extracted with 18 adjective - adjective seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

Note that manual inspection of the results showed that both resources did not include many good opposites, such as *bepaald - onbepaald* "determined - undetermined", *collectief - individueel* "collective - individual", *noordelijk - zuidelijk* "northern - southern" and others. Because of that and because the results in the study on strictly textual patterns showed that lexical resources do not provide reliable means for evaluation of the results, as the next step, all found pairs were also evaluated by participants. We asked three native speakers of Dutch to classify all pairs with scoring ≥0.9 as opposites or non-opposites. The results are presented in Table 5.4.

Participants achieved a high Fleiss's kappa score of 0.74. Recall that in a similar classification of pairs found with 18 adjective - adjective seeds with strictly textual patterns, participants achieved a Fleiss's kappa score of 0.66 (see Section 4.4.1.1, Chapter 4 for details). This is an interesting result, given that in the latter case, the participants had to evaluate 475 pairs, whereas in this study they evaluated 1,563 pairs. Thus, in this case they demonstrated a much higher level of agreement, despite the fact that they had to evaluate three times as many pairs. The main difference between the results is that this time all 1,563 found pairs were expressed by adjectives, whereas strictly textual patterns found pairs expressed not only by adjectives but also nouns. This shows that participants find it easier to decide whether a pair is antonymous or not for strictly adjectival pairs.

Out of 1,563 pairs, 33% were judged as opposites (517 pairs), 73.5% of which received unanimous vote (380 pairs). This means that our method identified 517 adjective - adjective opposites, a huge number in comparison to previous corpus-based work on antonymy (for example, with fewer than 30 adjective opposites studied in Jones [2002]). Among unanimously judged opposites were pairs *betaald - gratis* "paid - free", *machteloos - machtig* "powerless - powerful", *klassiek - nieuw* "classical - new".

Opposites by the majority vote included pairs that did not evoke a specific scale, for example, *illegaal - officieel* “illegal - official” or pairs that did not identify the opposite poles on it, for example, *primair - secundair* “primary - secondary”. Others belonged to non-binary sets and were not mutually exclusive although they indicated contrastive pairs, for example, *dierlijk - menselijk* “from animal - human”.

Among unanimously judged non-opposites (919 pairs in total) were pairs like *interessant - mooi* “interesting - nice”, *grijs - zwart* “grey - black”, *Albanees - Servisch* “Albanian - Serbian”. These pairs were found in very productive surface patterns of different pattern types. For example, the pair *interesting - nice* was found in patterns like [*tussen <ANT/ADJ> en <ANT/ADJ>* .] “between <ANT/ADJ> and <ANT/ADJ> .”, [*wat ik <ANT/ADJ> of <ANT/ADJ>*] “what I <ANT/ADJ> or <ANT/ADJ>”; the pair *grey - black* was found in patterns like [*, <ANT/ADJ> of <ANT/ADJ> ,*] “, <ANT/ADJ> or <ANT/ADJ> ,”, [*met <ANT/ADJ> en <ANT/ADJ>*] “with <ANT/-ADJ> and <ANT/ADJ>”; the pair *Albanian - Serbian* was found in patterns like [*de grens tussen <ANT/ADJ> en <ANT/ADJ>*] “the border between <ANT/ADJ> and <ANT/ADJ>”, [*, over <ANT/ADJ> en <ANT/ADJ>*] “, above <ANT/ADJ> and <ANT/ADJ>”.

These examples demonstrate that found pairs judged as non-opposites were often found in variations of the well-established ‘pattern of incompatibility’ [*between <ANT> and <ANT>*] (Lin et al. [2003]). Other patterns that were responsible for finding non-opposites were variations of rather general patterns like [*<ANT> or <ANT>*] that have also previously been discussed in relation to antonymy (Jones [2002], Jones et al. [2007]). In particular, Jones et al. [2007] argued that such patterns indicate antonym canonicity, so that more canonical opposites occur in a wider range of such patterns and fewer canonical opposites occur in a smaller number of such patterns. But what was missing in their analysis is a comparison to the corpus-behaviour of non-opposites. In other words, while Jones et al. [2007] studied a number of opposites, showing that indeed they occurred in such patterns, they did not conduct a comparative analysis of non-opposites to show that non-opposites do *not* co-occur in such patterns. Our results show that opposites as well as non-opposites (in the conventional sense) co-occur in such patterns. This implies that an explanation for antonym canonicity based on the breadth of co-occurrence is not enough to explain a strong association between opposites like *rich - poor*, *old - young* and others.

Another point highlighted by many pairs that were judged as non-opposites, is the weakness of the manual evaluation. Unexpectedly, pairs *klein - oud* “little - old”, *groen*

- *rood* “green - red”, and *licht* - *serieus* “licht - serious” were unanimously judged as non-opposites although in certain contexts these pairs are opposites. For example, *klein* “little” and *oud* “old” are opposites in the context of age, *groen* “green” and *rood* “red” in the context of ripeness as well as traffic lights, and *licht* “light” and *serieus* “serious” in the context of, for example, a degree of damage, or reading materials. It seems that it is difficult to recognize such pairs as opposites outside of their context.

In fact, our algorithm identified 13 pairs with the word *groen* “green” with scoring above 0.9. The pairs included:

rijp “ripe”,	*grijs* “grey”,	*paars* “purple”,
rood “red”,	*wit* “white”,	*ethisch* “ethical”,
blauw “blue”,	*zwart* “black”,	*duurzaam* “durable”,
geel “yellow”,	*bruin* “brown”,	*sociaal* “social”,
rose “pink”.		

Pairs *green* - *durable* and *green* - *social* were discarded as they did not co-occur with each other significantly often. All other pairs were judged as non-opposites by all three participants and none of them was listed as opposites in either CORNETTO or *MWB*. However, when the context of these pairs is taken into account, it becomes clear that “green” and “ripe” *are* opposites in the context of maturity; “green” and “black” are opposites in the context of coffee blends, as well as types of olives; “green” and “grey” are opposites when experienced and inexperienced people are compared; “green” and “yellow” are opposites when the ripeness of, for example, bananas is discussed, similar to “green” and “brown” in the context of the ageing of food. Finally, the pair “green” - “ethical” was found in the context of the comparisons between different types of investments (*groene of ethische beleggingen*), however, it does not seem to be contrastive.

The fact that participants failed to recognize such opposites in our evaluation task suggest that a better way of evaluation of opposites by participants might be the one that includes context, in which these pairs were found, such as, a noun modified by both adjectives. That is, instead of showing participants the pair *green* - *red* in isolation, it could be more helpful to show them complete noun phrases, such as, *a green apple* - *a red apple*. This has not been implemented in any of the existing evaluations because most of such tasks focused on well-established and therefore highly associated opposites.

Most importantly, what the aforementioned examples show is that a method based on a set of seeds and surface PoS patterns is capable of finding a wide range of opposites, canonical as well as non-typical, context-dependent pairs. The latter class is

Scoring level	All found pairs	Precision	Pairs with significant co-oc.	Precision
≥0.98	897	0.38	877	0.39
≥0.96 <0.98	222	0.25	207	0.26
≥0.94 <0.96	185	0.2	169	0.22
≥0.90 <0.94	337	0.19	310	0.2

Table 5.5: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 adjective - adjective seeds.

difficult for participants to think of on their own. Partially, this is why such pairs are missing in existing computational lexical resources such as CORNETTO.

Non-opposites that did not receive unanimous votes (11.2% or 131 pairs) also contained interesting pairs that should not necessarily be dismissed from the results. Some of them were very similar to those judged as opposites by the majority vote, for example, pairs *huidig - nieuw* "current - new", *huidig - toekomstig* "current - upcoming / future", *half - heel* "half - full", *gouden - zilveren* "golden - silver", *internationaal - Nederlands* "international - Dutch" (equal to *local* in Dutch newspaper texts). Other pairs were not strict opposites but they indicated contrastive concepts, for example *humanitair - militair* "humanitarian - military" in relation to different means of presence in other countries. Thus, pairs that are not unanimously judged as non-opposites are rather context-dependent contrastive pairs that would be useful for many NLP applications. Such pairs should not, therefore, be discarded from automatically identified results.

Unfortunately, although most of the pairs that did received the majority vote are contrastive in certain contexts, they were discarded from the results when we calculate the precision scores. However, all found pairs that were identified in the lexical resources as opposites were treated as unanimously judged opposites and, therefore, they were included when precision scores were calculated.

Thus, precision scores were calculated by taking into account pairs that were unanimously judged as opposites and pairs that were identified as opposites in the lexical resources as true positives and pairs that were unanimously judged as non-opposites as false positives. Pairs that did not receive unanimous votes were discarded. The results are summarized in Table 5.5.

Precision scores were low for all scoring levels, varying between 0.39 for 877 pairs

Top-k found pairs	Precision scores textual patterns	PoS patterns
50	0.88	0.61*
100	0.74	0.6*
150	0.6	0.57
200	0.54	0.53
250	0.5	0.5

Table 5.6: Top-k pairs with scoring $\geq$0.9 extracted with 18 adjective - adjective seeds. Precision scores are based on the classification of pairs by three participants. *based on the average of 10 randomly selected sets.

with significant co-occurrence and scoring $\geq$0.98 and 0.2 for 310 pairs with significant co-occurrence and scoring $\geq$0.90<0.94. As the results show, significant co-occurrence only slightly improved the precision.

Recall, however, that the precision scores based on scoring levels were also low for the results with textual patterns. For example, the precision score for 266 significantly co-occurring pairs with scoring $\geq$0.9 found with 18 adjective seeds was 0.49. Recall also that it seemed that the total number of found pairs played a role in that the precision scores were high when only top-k pairs were taking into consideration. For example, the precision score for the top-150 pairs found with textual patterns and 18 adjective seeds was 0.6. Then, it seems useful to compare the evaluation of the performance of textual and PoS patterns, by calculating the precision scores for top-k pairs. These scores are presented in Table 5.6.

Note, that while 18 adjective - adjective seeds found 40 pairs with the absolute score of one with textual patterns, they found 134 pairs with the score of one with PoS patterns. The same scoring implies that these 134 pairs are equally good candidate opposites, which makes it difficult to decide which 50 and 100 of them should be used for the assessment of the precision. As a result, we decided to compile ten sample sets of 50 and 100 pairs, calculate the precision scores for each of the sets and derive the average precision score based on the ten randomly selected samples.

At the first glance, the results seem to suggest that textual patterns perform better as they achieve higher precision scores for each set of top-k found pairs. For example, the precision score for the top-50 pairs from textual patterns is 0.88 and from PoS patterns - 0.61. While both scores are good, it is not clear why they differ by almost 34%. To understand this, we analysed the top-50 pairs found with PoS patterns (presented in Table 5.7), comparing them with sample sets of 50 pairs found (all with the score of

one) with textual patterns (one set is presented in Table 4.9).

As is shown in Table 5.7, PoS patterns found a wider range of opposites that belong to the same part-of-speech category. Among such pairs were typical opposites unanimously judged as such, for example, *passive - active*, as well as, non-traditional opposites that are not always recognized as such by judges, for example, *commercial - public*. For example, similar to the earlier mentioned pair *green - ripe*, the pair *white - red* was unanimously discarded by the judges, although these words are opposites in the context of wine types.

These pairs were not among top-50 pairs found with the same seed set in textual patterns. Instead, among top-50 pairs, textual patterns found the majority of the original seeds (89%) as well as readily recognized opposites expressed by nouns. In particular, 30% of top-50 pairs found with strictly textual patterns were not expressed by adjectives, which constituted 35 pairs, including 16 original seeds (45.7%). Of course, with PoS patterns, almost all pairs (96%) were expressed by adjectives (not 100% due to the pos-tagging errors). This result indicates that, although the precision scores with PoS patterns were lower than with strictly textual patterns, PoS patterns performed better as they were able to retrieve novel, less typical opposites that belong to the target part-of-speech category but which are also more difficult to classify based on manual classification.

This is reflected in the total number of pairs judged as opposites by the majority vote. Namely, 517 pairs found with PoS patterns were judged as opposites by the majority vote, whereas 208 pairs found with textual patterns were opposites according to the majority vote. Less than half of the 208 pairs were expressed by adjectives. Also, recall that textual patterns found the same pairs in the top results for seed sets expressed by different part-of-speech categories. These pairs tend to be well-recognized opposites. Because of that, there were more unanimously judged opposites and consequently higher precision score.

In fact, the overlap between top-200 pairs found with two types of patterns consists of 71 pairs, which is less than half. Among such pairs were the adjectives *klassiek - populair* “classical - popular”, *religieus - seculier* “religious - secular”, *kansarm - kansrijk* “underprivileged - promising”, *mannelijk - vrouwelijk* “male - female” and so on. The overlap is small because it included only those pairs, in which both words were expressed by adjectives.

Coming back to an earlier point about the differences between canonical and non-canonical opposites in terms of their breadth of co-occurrence, the fact that non-typical

Dutch	English	Opposites
zwak - sterk	weak - strong	yes unanimously
Amerikaans - Nederlands	American - Dutch	no unanimously
koud - warm*	cold - hot*	yes unanimously
kort - middellang	short - middle long	no by majority vote
commercieel - publiek	commercial - public	no by majority vote
intern - extern	internal - external	yes unanimously
nieuw - oud*	new - old*	yes unanimously
gelukkig - ongelukkig	happy - unhappy	yes unanimously
wit - rood	white - red	no unanimously
dom - slim	stupid - smart	yes unanimously
ziek - oud	sick - old	no unanimously
westers - islamitisch	western - Islamic	no unanimously
langzaam - snel*	slow - fast*	yes unanimously
abstract - figuratief	abstract - figurative	yes unanimously
etnisch - religieus	ethnic - religious	no unanimously
mannelijk - vrouwelijk	male - female	yes unanimously
westers - oosters	western - eastern	yes unanimously
zacht - hard*	soft - hard*	yes unanimously
Vlaams - Nederlands	Flemish - Dutch	no unanimously
groen - rijp	green - ripe	no unanimously
analoog - digitaal	analogue - digital	yes unanimously
geheel - gedeeltelijk	complete - partial	yes unanimously
mooi - goed	nice - good	no unanimously
modern - traditioneel	modern - traditional	yes by majority vote
jong - oud*	young - old*	yes unanimously
lichamelijk - geestelijk	bodily - mental	yes unanimously
blind - slechtziend	blind - with poor eyesight	no by majority vote
laag - hoog*	low - high*	yes unanimously
passief - actief*	passive - active*	yes unanimously
bekend - onbekend	known - unknown	yes unanimously
positief - negatief	positive - negative	yes unanimously
gelovig - ongelovig	religious - non-religious	yes unanimously
maatschappelijk - politiek	social - political	no unanimously
chemisch - biologisch	chemical - biological	no unanimously
zwaar - licht*	heavy - light*	yes unanimously
niet-werk - werk	not-work - work	yes by majority vote
economisch - politiek	economical - political	no unanimously
heel - half	complete - half	no by majority vote
bijzonder - openbaar	exceptional - public	no unanimously
Nederlands - Duits	Dutch - German	no unanimously
militair - politiek	military - political	no unanimously
onbewust - bewust	unaware - aware	yes unanimously
protestants - katholiek	Protestant - Catholic	no by majority vote
onecht - echt	false - real	yes unanimously
internationaal - nationaal	international - national	yes unanimously
conservatief - progressief	conservative - progressive	yes unanimously
arm - rijk*	poor - rich*	yes unanimously
koud - heet	cold - hot	yes unanimously
allochtoon - autochtoon	foreign - indigenous	yes unanimously
vrouwen - mannen	women - men	yes unanimously

Table 5.7: A sample of fifty pairs with the score of one found with 18 adjective - adjective seeds by means of PoS patterns and their classification according to three judges.

Top-k found pairs	Pairs found with 6 seeds	12 seeds	18 seeds
50	**0.77**	0.7	0.61*
100	0.61	**0.66**	0.6*
150	0.56	**0.57**	**0.57**
200	**0.55**	0.53	0.53
250	**0.53**	0.49	0.5

Table 5.8: Top-k pairs with scoring ≥0.9 extracted with six, 12 and 18 adjective - adjective seeds. Precision scores are based on the classification of pairs by three participants and lexical resources.

pairs like *white - red* and *conservative - progressive* were found with very productive PoS patterns, achieving the highest possible scoring, suggests that significant co-occurrence in patterns of incompatibility is a property of non-canonical pairs as much as it is of canonical opposites. We mentioned earlier that Jones et al. [2007] examined the range of patterns, in their terms, the breadth of co-occurrence, only in relation to canonical opposites, neglecting non-opposites. These examples show that they have also neglected to study the behaviour of non-canonical opposites in comparison to canonical ones. Non-canonical pairs were not taken into account under an implicit assumption that unlike canonical opposites, they also do not co-occur in a wide range of patterns (similarly to the assumption that non-opposites do not co-occur in a wide range of patterns). Our results suggest that there is no ground to support that assumption because we show that both canonical *and* non-canonical opposites demonstrate a similar pattern in behaviour: they co-occur in the same patterns equally often. This further implies that neither significant co-occurrence nor the breadth of co-occurrence are sufficient for identification of canonical opposites.

The next question that needs to be addressed is how the number of seeds affects the results. Recall that at the beginning of this section we said that a set of six seeds found fewer pairs (635) than sets of 12 (2,595) and 18 seeds (3,275 pairs). The difference in the number of found seeds also remained for the pairs with the highest scoring. Namely, at the score level of ≥0.9, the six seeds found 297 pairs while the set of 18 seeds - 1,641 pairs. We have already shown that the best opposites are among the top 250 found pairs. Given that the set of 18 seeds found all pairs found by the set of six and 12 seeds, what is the possibility that the best opposites are among 297 pairs with the score ≥0.9 that were already found with the set of six seeds?

The precision scores for top-k pairs found with seed sets of six, 12 and 18 pairs of

opposites are given in Table 5.8. Due to the larger number of original seeds, the set of top-50 pairs found with six seeds has a higher precision score (namely, 0.77) than the sets of pairs found with 12 and 18 seeds (0.7 and 0.61 respectively). As the number of found pairs increases to 100, all three sets perform well gaining a similar precision between 0.61 (six seeds) and 0.66 (12 seeds). Up to 150 pairs all three seed sets find a similar number of opposites. Interestingly, after that a set of six seeds performs better than larger sets of 12 and 18 seeds. For the top-250 pairs, the precision score for pairs found with six seeds was as high as 0.53, for pairs found with 12 seeds it is 0.49 and for pairs found with 18 seeds, it was 0.5. This is an encouraging result as it shows that the program will perform equally well even with a very small set of adjectival seeds. Since the program takes less time to run (up to five days), this means that fewer seeds can be more efficient when less computational power is available or when a larger corpus is used.

Finally, we can use manual classification of found pairs to evaluate the coverage of opposites expressed by adjectives in CORNETTO. In particular, we can examine how many pairs that were judged as opposites by the majority vote (at least two participants) were present in this resource and how many of them were linked as opposites. Out of 517 opposites according to the majority vote, 98.8% (511 pairs) had both words present in CORNETTO and 19% of them (97 pairs) were marked as opposites. Among missing opposites were pairs *tijdelijk* - *vast* "temporary - permanent", *gesponsord* - *niet-gesponsord* "sponsored - not-sponsored", *eerlijk* - *oneerlijk* "honest - dishonest", *koel* - *warm* "cool - warm" and other opposites.

Out of 249 opposites identified by means of CORNETTO and *MWB*, 209 pairs were among 517 opposites according to manual evaluation. Thus, the algorithm found 84% of opposites represented in this computational lexical resource among identified pairs. In addition, our method found another 308 pairs (59.6%) which were recognized as opposites by the participants but are not currently represented in CORNETTO.

5.4.1.2 Acquired patterns with part-of-speech information

A total of 18,983 unique patterns with part-of-speech information were acquired with the set of 18 adjective - adjective seeds. The shortest patterns were four tokens long (13.5%), and the longest patterns were seven tokens long (19.9%). Thus, although it was possible to have patterns that consist of three tokens only, such patterns were discarded, most likely because they were too general. Also, the fact that the patterns

at higher scoring levels (>0.5) on average were longer (six tokens long) than patterns with lower scoring suggests that specific patterns were better at finding opposites.

Recall that the same seed set found more than 30,000 strictly textual patterns. This is an unexpected result, given that PoS patterns found more pairs than strictly textual patterns. The most productive pattern types of found PoS patterns were the same as the ones identified within strictly textual patterns, with the pattern type [*between* <*ANT*> *and* <*ANT*>] being by far the most productive pattern type. What this suggests is that when the part-of-speech category of found pairs is limited to adjectives, adjective - adjective seeds are able to identify the same pattern types among PoS patterns as among strictly textual patterns but the range of variations within the pattern types is smaller. In comparison to the results for strictly textual patterns, these results in a smaller number of total PoS patterns identified by the algorithm but a larger number of found pairs expressed by the same part-of-speech category with higher automatic scores.

5.4.2 Results for noun - noun seed pairs

Using a full version of TwNC, 3,941 pairs with frequency ≥5 were found with the set of six seeds, 4,025 pairs with the set of 12 seeds and 5,014 pairs with the set of 18 seeds. The overview of how many pairs were found with each set is given in Table 5.9. In general, seed sets expressed by nouns led to extraction of the largest number of pairs across all seed sets with the part-of-speech patterns. The same seed sets found fewer pairs also with strictly textual patterns. In particular, 18 noun - noun seeds with strictly textual patterns found 2,019 pairs with frequency ≥5, which is less than the set of six noun - noun seeds with PoS patterns. This shows that the PoS patterns are much more productive than strictly textual patterns.

Since all pairs found with six and 12 seeds were also found with 18 seeds, the results from the set of 18 seeds will be discussed next. Note that only pairs with scoring >0.6 were regarded as candidate opposites.

5.4.2.1 Found pairs

As is shown in Table 5.10, 94.4% of 4,805 pairs found with the set of 18 noun - noun seeds co-occurred with each other within a sentence in the newspaper corpus significantly more often than would be expected by chance. Out of 4,534 found pairs with significant co-occurrence, many were not opposites, especially among pairs with lower

Scoring	Pairs found with		
	6 seeds	12 seeds	18 seeds
≥0.9	42.6% (1,681)	42.7% (1,720)	44% (2,205)
≥0.8<0.9	20.2% (794)	20.4% (819)	20% (1,003)
≥0.7<0.8	18.5% (729)	18.3% (736)	18.4% (922)
≥0.6<0.7	13.8% (545)	13.9% (559)	13.5% (675)
<0.6	4.9% (192)	4.7% (191)	4.2% (209)
Total	*3,941*	*4,025*	*5,014*

Table 5.9: Total number of pairs found with six, 12 and 18 noun - noun seeds in the TwNC by means of surface patterns with PoS information (only with frequency >5).

Scoring	Number of pairs	Significant co-occurrence
≥0.9	2,205	96.7% (2,132)
≥0.8<0.9	1,003	93.1% (931)
≥0.7<0.8	922	91.7% (844)
≥0.6<0.7	675	93.2% (627)
Total	*4,805*	94.4% (4,534)

Table 5.10: Number of unique pairs found by means of PoS patterns with 18 noun - noun seeds in the TwNC per scoring level, and number of pairs that co-occurred with each other sententially significantly more often than would be expected by chance.

scoring level. This shows that significant co-occurrence is particularly weak at predicting antonymy among candidate pairs expressed by nouns. It might be that significant co-occurrence is not sufficient for separating opposites from non-opposites because many noun - noun pairs tend to co-occur with each other significantly often. However, as will be discussed later, significant co-occurrence and automatic scoring (that is the highest scoring of one) together can be used to reliably identify opposites. Because manual inspection showed that all found opposites were among top-found results based on the automatic scoring, pairs with the score <0.9 were dismissed from the results.

Out of 2,132 pairs with significant co-occurrence and score ≥0.9, 86.8% (1,851 pairs) had both words present in CORNETTO but only 1.9% of them (36 pairs) were linked as opposites. Among identified opposites were pairs *winnaar - verliezer* "winner - loser", *meneer - mevrouw* "Mister - Mrs.", *export - import* "export - import" and others. Seventy-fine percent of opposites were linked with each other symmetrically, for example, the pair *werknemer - werkgever* "employee - employer", but in 25% of

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.98	1,176	2.8% (29/1,033)	5.8% (69)	6.6% (78)
≥0.96<0.98	291	0 (0/247)	1.4% (4)	1.4% (4)
≥0.94<0.96	231	0.5% (1/199)	2.6% (6)	2.6% (6)
≥0.90<0.94	434	1.6% (6/372)	2.3% (10)	3.4% (15)
Total	*2,132*	1.9% (36/1,851)	4.2% (89)	4.8% (103)

Table 5.11: Pairs with scoring ≥0.9 found with 18 noun - noun seeds in TwNC by means of PoS patterns and the number of pairs that were found in one or both of the lexical resources (CORNETTO and *Mijnwoordenboek.nl* (*MWB*)).

pairs only one of the words was listed as an opposite of the other, for example, the pair *koper - verkoper* "buyer - seller". Again, this highlights that opposition is covered in CORNETTO inconsistently.

Recall that more than half (71%) of all adjectival opposites identified by one or both of the resources were present in CORNETTO. Also *MWB* contained 67% of all opposites, identified as such in one or both of the resources (224 opposites in total). In the case with noun - noun pairs, out of 103 opposites identified by means of lexical resources, most of the opposites were identified with the help of *MWB* (89 pairs or 86.4%) rather than CORNETTO (36 pairs or 35%). One of the reasons for this can be the fact that opposites expressed by nouns are not well covered in CORNETTO. For example, among opposites that were not listed as such in this lexical resource were pairs *leugen - waarheid* "lie - truth", *haat - liefde* "hatred - love", *armoede - rijkdom* "poverty - wealth" and so on. This illustrates how automatically extracted opposites can improve the coverage of antonymy even in the most-up-to date computational lexical resources like CORNETTO.

When both resources were used for identification of opposites, a mere 4.8% (103) of pairs were opposites among 2,132 pairs with the score ≥0.9. In comparison, using the same set of noun seeds, strictly textual patterns found 960 pairs with scoring ≥0.9, 8.7% (84) of which were opposites according to the lexical resources. Given that the same corpus and the same set of seeds were used, this seems to show that PoS patterns extract many more pairs than strictly textual patters without improving the precision. Even when only pairs with scoring ≥0.98 (1,198 pairs) are taken into consideration, strictly textual patterns lead to higher precision (84 opposites as opposed to 78 pairs). This is a preliminary conclusion, however, and as we will show below, the assumption that strictly textual patterns perform better is not correct. We will show that the main

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	70.9% (283)	68.2% (193)	51.5% (893)	83.3% (744)	1,176
≥0.96<0.98	7.3% (29)	48.3% (14)	15.1% (262)	87.8% (230)	291
≥0.94<0.96	9.8% (39)	64.1% (25)	11.1% (192)	76.6% (147)	231
≥0.90<0.94	12% (48)	68.7% (33)	22.3% (386)	88.6% (342)	434
Total	*18.7% (399)*		*81.3% (1,733)*		*2,132*

Table 5.12: Percentage of pairs with scoring ≥0.9 extracted with 18 noun - noun seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

reason why there were more opposites present among the set of pairs found by means of textual patterns is because it contained not only noun - noun but also adjective - adjective pairs, including adjective seeds, which are better covered in CORNETTO. In other words, the set of 84 opposites found by means of strictly textual patterns contain not only noun - noun pairs, but the most frequent pairs, including canonical adjectival pairs like *young - old*. On the other hand, the set of 78 pairs identified as opposites by lexical resources found by means of PoS patterns contain only noun - noun pairs. Therefore, PoS patterns find more opposites expressed by nouns.

Given that so many opposites expressed by nouns seem to be missing from CORNETTO, it is particularly interesting to know how many found pairs were judged as opposites by participants. We asked three participants to classify all pairs with the score ≥0.9 as opposites or non-opposites. Participants achieved a Fleiss's kappa score of 0.617 which indicates substantial agreement. Nevertheless, participants demonstrated a stronger inter-annotator agreement when classifying adjective - adjective pairs (Fleiss's kappa score of 0.74), suggesting that classification of pairs expressed by nouns was a more difficult task. The results of manual classification are summarized in Table 5.12.

Out of 2,132 candidate pairs, 18.7% were judged as opposites (399 pairs) and 81.3% were judged as non-opposites (1,733 pairs). More than 70% of pairs that were judged as opposites had a score above 0.98 (283 pairs in total). Among unanimously judged opposites, which made up 66.4% of all opposites, were pairs *vinder - zoeker* "finder - seeker", *dieptepunt - hoogtepunt* "low-point - peak", *huur - koop* "rent - purchase", *regen - zon* "rain - sun".

Almost 85% of pairs judged by the participants as non-opposites received unanimous votes. Among such pairs were frequently co-occurring pairs like *computer - internet* "computer - Internet", *verkeer - weg* "transport - road", co-hyponyms like

brandweer - politie “fire department - police”, *auto - fiets* “car - bicycle”, *cello - viool* “cello - violin” and words that co-occur in fixed expressions, for example, *schip - wal* in *‘De wal zal het schip keren’*, meaning that the course of things will take a different turn automatically (literary translation “The shore will stop the ship”). This shows that, similar to strictly textual patterns, surface PoS patterns also frequently contain semantically similar words and words from fixed expressions. This highlights that significant co-occurrence as an additional constraint does not suffice for eliminating non-antonymic frequently co-occurring words from the results in any pattern-based method for finding opposites.

However, from a linguistic point of view, the two most interesting groups of found pairs are opposites and non-opposites that did *not* receive unanimous votes because these are the pairs that caused disagreement among the participants and were not unanimously discarded as non-opposites. For example, the pair *dier - mens* “animal - human” was judged by the majority vote as opposites while the pairs *ding - mens* “thing - human” and *computer - mens* “computer - human” were judged by the majority vote as non-opposites, although in both pairs words are contrasted on the animacy scale (as animate - non-animate). Following theoretical approaches, in particular Murphy [2003], similar to mutual incompatibles, pairs like *human - computer* bisect some domain and, as a result, an assertion *X is a computer* contradicts *X is a human* and vice versa. Therefore, such pairs are contradictory opposites. But why is it that participants did not recognize them as opposites?

Pattern analysis shows that *animal - human*, *thing - human* and *computer - human* were found in different variations of the same pattern types: [*between* <*ANT*> *and* <*ANT*>], [<*ANT*> *as well as* <*ANT*>], [<*ANT*> *and* <*ANT*> *alike*], [*from* <*ANT*> *to* <*ANT*>]. The pairs differed as to their overall frequency of co-occurrence: the pair *computer - human* was found 36 times, the pair *thing - human* was found 62 times and the pair *animal - human* was found 728 times. Further, while pairs *computer - human* and *thing - human* were found in the contexts in which their differences were emphasized, the pair *animal - human* often occurred in the contexts in which the similarities, especially equality, between the two were emphasized. For example, the pairs *thing - human* and *computer - human* were frequently found within contrastive patterns of the type [*onderscheid / verschil tussen* <*ANT*> *en* <*ANT*>] “difference between <ANT> and <ANT>” while the pair *animal - human* frequently co-occurred in patterns of the type [*de gelijkheid / gelijkstelling / gelijk recht van* <*ANT*> *en* <*ANT*>] “equality / equalization / equal right of <ANT> and <ANT>”. This suggests that

there is an inherited underlying difference between animals and humans perceived by participants, which is not present when humans are compared to inanimate objects. As a result, although the latter co-occur in contrastive contexts, they are not perceived as contradictory opposites. Rather they express a more pragmatic contrast which is not recognized by the participants outside of the context.

- **mens - computer**: Is er dan echt geen verschil tussen *mens* en *computer*?
- *human - computer*: Is there then really no difference between *human beings* and *computers*?
- **mens - dier**: 'Hoewel het in eerste instantie lijkt alsof dierrechtorganisaties streven naar de gelijkheid van *mens* en *dier*', zegt Parmentier in De Groene Amsterdammer, 'worden uiteindelijk de mens rechten ontnemen: het gebruik van dieren.'
- *human - animal*: 'No matter how much it seems at first glance as if organizations for animal rights are fighting for the equality of humans and animals', says Parmentier in "De Groener Amsterdammer", 'at the end the rights of people are taken away: the right to use animals.'

Similar to *thing - human* and *computer - human*, the pair *mens - machine* "human - machine" co-occurred 68 times. It bisects a domain, so *X is a machine* entails *X is not a human being*. This pair was found in the variations of two types of patterns [*difference/relationship between <ANT> and <ANT>*] and [*either <ANT> or <ANT>*]. But, although its behavioural profile in the corpus was similar to the pairs that were judged as non-opposites by the majority vote, this pair was unanimously judged as an opposite. One possible reason for this result is that since all participants were students at the Faculty of Natural Sciences, they recognized the pragmatic contrast between *human - machine* more readily. If this is the case, this example illustrates several important points.

Recall that human intuition is particularly weak when it comes to non-typical opposites like *red - white* (wine), especially in the absence of context. So far, corpus-based studies on antonymy have focused exclusively on differences between canonical pairs like *rich - poor* and their non-canonical counterparts like *rich - wealthy* because it seems necessary to understand the nature of canonicity on the example of readily recognized opposites before studying the differences between non-canonical opposites like

red - white and contrastive pairs like *human - machine* that require additional context. Our results, however, show that all these pairs exhibit similar behaviour in the corpus, showing that in natural language production they similarly co-occur in the same contrastive patterns. Previous theoretical accounts often discard such 'difficult' cases, following researcher's intuition about canonicity of opposites. Corpus evidence, on the other hand, provides a more structured and reliable means of analysis and must be taken into account.

Another pair that was not recognized as antonymous outside of the context was *vis - vlees* "fish - meat", which was judged as non-opposites by the majority vote. Among patterns that found the pair were variations of pattern types [<*ANT*> *or* <*ANT*>], [<*ANT*> *as well as* <*ANT*>], [*more / less* <*ANT*> *than* <*ANT*>], [*between* <*ANT*> *and* <*ANT*>] as well as [*be neither* <*ANT*> *nor* <*ANT*>]. The latter pattern could also be part of a fixed expression *vlees noch vis*, meaning "neither fish nor fowl". But because this pair was found in this fixed expression only four times, out of 32 occurrences, it is safe to conclude that *fish - meat* tend to co-occur outside of the fixed expressions in most of the co-occurrences. For example, this pair was found in the following sentence:

- **vis - vlees**: Toch zijn we geen uitgesproken viseters; we hebben een keurig balans tussen *vis* en *vlees*.
- *fish - meat*: At the end we are not such serious eaters of fish; we have a sensible balance between *fish* and *meat*.

Relying solely on patterns can be misleading. For example, the pair *vlees - bloed* "flesh - blood" was found 49 times, always as part of the idiomatic expression "flesh and blood", meaning 'human nature'. Also the pair *vlees - melk* "meat - milk" was found 25 times in the variations of pattern types [*division between X and Y*], [<*ANT*> *or* <*ANT*>] and [<*ANT*> *as well as* <*ANT*>]. These words were found in two contexts: eating regulations related to kosher food and milk and meat production by *'dubbeldoel'-koeien* or "all-purpose" cows, that is fast growing cows that produce a lot of milk and meat. Although they pass the significant co-occurrence test and they are found in patterns that tend to contain opposites, they are not antonymous.

Out of 103 opposites identified in the lexical resources and 399 opposites according to the majority vote, 95 pairs were the same, which shows that 92% of noun - noun opposites listed in CORNETTO can be found in the 350 million words corpus of newspaper

Scoring level	All found of pairs	Precision	Pairs with significant co-oc.	Precision
≥0.98	1,198	0.21	1,176	0.22
≥0.96 <0.98	304	0.06	291	0.06
≥0.94 <0.96	244	0.14	231	0.14
≥0.90 <0.94	459	0.09	434	0.1

Table 5.13: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 noun - noun seeds.

texts, a relatively small corpus in comparison to studies that use the Web for finding lexical relations like hyponyms and meronyms (for example, Pantel and Pennacchiotti [2006]).

In order to assess precision scores, we combined found opposites that were listed in the lexical resources together with found pairs that were unanimously classified by the participants as opposites. These pairs were used as true positives. All found pairs that were unanimously judged as non-opposites were used as false positives. All pairs that did not receive unanimous votes were discarded.

The precision scores for all found pairs based on their automatic scoring are presented in Table 5.13. As can be seen, the precision is very low even for the pairs at the highest score level ≥0.98. Significant co-occurrence improved the precision only slightly. For example, at the score level ≥0.98, the precision score for all 1,198 found pairs is 0.21 and for 1,176 pairs that co-occurred with each other significantly often, the precision score is 0.22. However, as previous results have shown, it is more useful to assess precision scores for the top-k found pairs, as the number of possible opposites that can be found automatically seems to be limited to the top-k pairs, in which the number k depends on the part-of-speech category of found opposites.

According to the results summarized in Table 5.14, the precision score for the top-50 pairs found by means of surface PoS patterns is 0.63. This is lower than the precision score for the top-50 pairs found by means of strictly textual patterns. There are two reasons why textual patterns outperformed PoS patterns for the top-50 pairs. First, there was a difference between the sets of found pairs in relation to the number of the original seeds they contained. In particular, the precision score of 0.74 is based on one set of 50 pairs found with textual patterns. This set included 12 of 18 original noun - noun seeds, which made up 24% of the total number of pairs used for calculation of

Top-k found pairs	Precision scores for textual patterns	Precision scores for PoS patterns
50	**0.74**	0.63*
100	0.59	**0.61***
150	0.5	**0.56**
200	0.44	**0.48**
250	0.42	**0.43**

Table 5.14: Top-k pairs with scoring $\geq$0.9 extracted with 18 noun - noun seeds. Precision scores are based on the classification of pairs by three participants. *based on the average of ten randomly selected sets of pairs that had the highest automatic scoring of one.

the precision. The precision score of 0.63 for PoS patterns is based on the average of 10 precision scores calculated for 10 randomly selected sets of found pairs with the highest automatic scoring of one. This was done because more than 100 pairs found with 18 noun - noun seeds in PoS patterns received a score of one, which means that they are all equally good candidate pairs. As a result, it was possible to create multiple sets of top-50 pairs. And, although the original seeds were also present, they did not influence the precision score as much as in the case of textual patters. This was not possible to do with the results for textual patterns because only 43 found pairs had an automatic score of one and 12 original seeds were among them.

The second reason why textual patterns had a higher precision score for top-50 found pairs than PoS patterns is because they found different types of pairs. While textual patterns found other well-established opposites expressed by other part-of-speech categories, particularly adjectives, including original adjective-adjective seeds, surface PoS patterns found a wider range of pairs expressed by nouns only. As a result, such pairs caused more disagreement among judges and the precision score is lower. But, as has been discussed in Chapter 4, strictly textual patterns tend to find a small number of *same* opposites across seed sets of different syntactic categories. Surface PoS patterns, on the other hand, find a wider range of opposites of the same syntactic category as the seed set. This difference is already reflected in the precision scores for top-100 pairs, in which PoS patterns achieve a higher precision score than (namely, 0.61) than textual patterns (namely, 0.59).

Because of the ability of surface PoS patterns to find many opposites expressed by nouns, the results of such methods are particularly useful for improvement of the existing computational lexical resource CORNETTO. For example, out of 399 opposites

by the majority vote that were found by means of PoS patters, 316 pairs (79%) had both words present in CORNETTO but only 33 (10.4%) of them were marked as opposites. This means that for 283 found noun - noun opposites, which constitute 89.5% of identified opposites that have both words present in CORNETTO, the relationship of antonymy is not indicated. These opposites are expressed only by nouns, which means that a pattern-based method that restricts the syntactic category of candidate opposites is particularly useful for finding many opposites of the same part-of-speech category.

In contrast, out of 232 opposites by the majority vote that were found with strictly textual patterns, 199 pairs (86%) were present in CORNETTO but only 37 of them (18.6%) were marked as opposites. Thus, antonymy relationship was not indicated between 162 opposites found by means of strictly textual patterns. This constitutes 69.8% of automatically identified opposites that already have both words present in CORNETTO. In comparison to this result, PoS patterns identified 121 more opposites, all of which already have both words present in CORNETTO but not linked as opposites. All these pairs are expressed by nouns whereas opposites found by means of textual patterns are expressed not only by nouns but also by adjectives and verbs.

Note that PoS patterns also found derivations of seeds from other syntactic categories among top found pairs. This is an interesting result as it sheds light on antonym canonicity from a novel perspective. Recall that previous accounts on canonicity have linked it to the tendency of opposites to co-occur with each other significantly often in a wide range of different patterns. Since surface PoS patterns found derivations of canonical adjectival seeds, it seems that antonym canonicity also manifests itself in the possibility for well-established opposites, or rather, concepts they refer to, to be expressed by other syntactic categories. As a result, a pattern-based method can find them automatically.

Table 5.15 gives an example of what kind of pairs were found by means of 18 noun - noun seeds. It contains a sample of 50 randomly selected pairs that received the score of one. In this sample, among unanimously judged opposites were pairs *jongen - meisje* “boy - girl”, *regel - uitzondering* “rule - exception”, *dochter - zoon* “daughter - son”, *docent - leerling* “instructor - pupil”, *kind - ouder* “child - parent”, *democraat - republikein* “Democrat - Republican” and so on. According to the existing classifications, some of these examples are gender opposites, for example, *boy - girl*; some are converse opposites, for example, *child - parent*; and some are incompatibles, for example, *Democrat - Republican*. It is more difficult to apply existing classifications to other pairs, like *friend - family*, *luck - wisdom*, and so on as such examples do not

Dutch	English	Opposites
kerk - staat	church - state	yes unanimously
homoseksueel - vrouw	homosexual - woman	no unanimously
katholiek - protestant	Catholic - Protestant	no by majority vote
parlement - regering	parliament - government	no unanimously
chaos - orde	chaos - order	yes unanimously
homo - lesbiennes	homo - lesbian	yes unanimously
jongen - meisje	boy - girl	yes unanimously
gemeente - provincie	municipality - province	no unanimously
norm - waarde	norm - value	no by majority vote
feit - fictie	fact - fiction	yes by majority vote
familie - vriend	family - friend	yes by majority vote
minister - staatssecretaris	minister - State Secretary	no unanimously
regel - uitzondering	rule - exception	yes unanimously
Democraat - Republikein	Democrat - Republican	yes unanimously
hogeschool - universiteit	college - university	no unanimously
vakbeweging - werkgever	trade union - employer	yes by majority vote
docent - leerling	instructor - pupil	yes unanimously
dochter - zoon	daughter - son	yes unanimously
burger - overheid	citizen - authorities	no by majority vote
leerling - leraar	pupil - teacher	yes unanimously
moslim - niet-moslim	Muslim - not-Muslim	yes by majority vote
jaar - maand	year - month	no unanimously
optimist - pessimist	optimist - pessimist	yes unanimously
bond - werkgever	union - employer	yes by majority vote
moeder - vader	mother - father	yes unanimously
vakbond - werkgeverorganisatie	trade union - employer organization	yes unanimously
arm - rijk	poor - rich	yes unanimously
hond - kat	dog - cat	no unanimously
dood - leven	death - life	yes unanimously
justitie - politie	Ministry of Justice - police	no unanimously
kind - vrouw	child - woman	no unanimously
vakbond - werkgever	trade union - employer	yes by majority vote
kind - ouder	child - parent	yes unanimously
geluk - wijsheid	luck - wisdom	no by majority vote
oorlog - vrede	war - peace	yes unanimously
kind - moeder	child - mother	no unanimously
prins - prinses	prince - princess	yes unanimously
dader - slachtoffer	offender - victim	yes unanimously
werkgever - werknemer	employer - employee	yes unanimously
hetero - homo	hetero - homo	yes by majority vote
kwaliteit - prijs	quality - price	no unanimously
land - stad	country (side) - city	no unanimously
heer - meester	mister - master	no unanimously
vijand - vriend	enemy - friend	yes unanimously
christen - jood	Christian - Jew	no unanimously
jood - moslim	Jew - Muslim	no unanimously
verlies - winst	loss - victory	yes unanimously
dochter - moeder	daughter - mother	yes by majority
fictie - werkelijkheid	fiction - reality	yes unanimously
Belg - Nederlander	Belgian - Dutchman	no unanimously

Table 5.15: A sample of fifty pairs with the score of one found with 18 noun - noun seeds by means of PoS patterns and their classification according to three judges.

clearly fall under any of the well-established categories.

The most frequent pair with the highest scoring was the pair *state - church*. It was found with patterns that emphasize differences:

- **kerk - staat**: Wel lijkt in Mexico de honderd jaar oud vijandschap tussen *kerk* en *staat* definitief ten einde, toen president Fox onder het oog van zijn onderdaan nederigen een kus heeft gedrukt op de 'visserring' aan de hand van de paus.
- *church - state*: It seems that the hundred years old animosity between the *church* and the *state* in Mexico has definitely ended, when president Fox kissed the 'fisherman's' ring on the hand of the Pope under the eyes of his humbled citizens.

or, on the contrary, diminish presupposed incompatibility between the two:

- **kerk - staat**: 'Er is', zegt Suk, 'voor de christen geen verschil tussen *kerk* en *staat*, tussen zakelijk en privé.'
- *church - state*: According to Suk, there is no difference for the Christians between the *church* and the *state*, between business and private.

In both cases, there seems to be a presupposed incompatibility between the two, which was also recognized by the participants who unanimously classified the pair as antonymous even without any context.

The pattern [*vijandschap tussen <ANT> en <ANT>*] "animosity between <ANT> and <ANT>" is itself a very strong indicator of incompatibility. It provides a useful tool for finding and analysing pairs that are perceived as non-contrastive co-hyponyms outside of the context. For example, this pattern found the following pairs:

- *de vijandschap tussen Frankrijk en Duitsland* ('the animosity between France and Germany')
- *tweeduizend jaar vijandschap tussen christen en joden* ('two thousand year animosity between Christians and Jewish')
- *vijandschap tussen Serviers en niet-Slaven (lees: het Westen)* ('animosity between Serbians and non-Slavs (read: the West)')
- *de vijandschap tussen de Palestijn en Israel* ('the animosity between the Palestine and Israel')

Top-k found pairs	Pairs found with 6 seeds	Pairs found with 12 seeds	Pairs found with 18 seeds
50	0.66*	**0.67***	0.63*
100	0.64	**0.66**	0.61*
150	0.52	0.53	**0.56**
200	0.43	0.45	**0.48**
250	0.38	0.39	**0.43**

Table 5.16: Top-k pairs with scoring $\geq$0.9 extracted with six, 12 and 18 noun - noun seeds. Precision scores are based on the classification of pairs by three participants and lexical resources. *based on the average of ten sample sets.

- *vermeende vijandschap tussen premier en president* ('alleged animosity between premier and president')

As can be seen from the examples above, this pattern type is particularly good at finding pairs that indicate, often contextual, opposition.

As other examples in Table 5.15 show, the participants also failed to recognize some of the established opposites, such as *country side - city*. This shows that the line between context-dependent and non-contextual opposites is rather blurred.

Next we investigate how the size of the seed set played a role in the results by comparing precision scores for top-k pairs found with six, 12 and 18 seeds. The results are presented in Table 5.16. Because more than 50 pairs found with the sets of six and 12 seeds had an automatic score of 1, ten sets of randomly selected 50 pairs with the highest score were created and the precision scores are reported as the average over the ten sets.

The precision scores were for the top-50 and top-100 found pairs were high for all three seed sets, ranging from 0.63 and 0.67 for the top-50 pairs and between 0.61 and 0.66 for the top-100 pairs. For up to the top-100 pairs, the set of 12 seeds outperformed the set of six and the set of 18 seeds.

To understand why the set of 12 seeds achieved higher precision scores than 18 seeds, we compared the overlap between pairs with the automatic scoring of one. While all but one pair, namely, *dead - wounded*, that were found in the set of 12 seeds were also found with the set of 18 seeds, the set of 18 seeds contained 34 more pairs, out of which 30 pairs were also found with the set of 12 seeds, but they did not receive the score of 1 and four were found only in the set of 18 seeds. It seems then that the set of 18 seeds finds more different pairs that have higher automatic scoring but do not always receive unanimous votes. Thus, depending on the goal of the task, more seeds

should be used when studying a wide range of typical and non-typical, or rather less intuitive and more context-dependent opposites, whereas fewer seeds should be used when studying more typical opposites.

Finally, to evaluate the coverage of noun - noun opposites in CORNETTO, we checked how many of found pairs that were judged as opposites by the majority vote were present in CORNETTO and how many of them were marked as opposites. Out of 399 opposites according to the majority vote, 349 (87.5%) were found in CORNETTO but only 33 of them (9.4%) were opposites. This confirms the suggestions made earlier, namely, that opposites expressed by nouns are not well represented in this computational lexical resource and automatic methods for finding opposites can be successfully used to expand the coverage of the antonymy relation for pairs already present in CORNETTO as well as novel opposites, especially domain-specific ones.

5.4.2.2 Acquired patterns with part-of-speech information

A total of 65,867 unique patterns with part-of-speech information were acquired with the set of 18 noun - noun seeds. The shortest patterns were four tokens long (4.6%), and the longest patterns were seven tokens long (30.7%). Thus, as with PoS patterns found by means of adjectives, as well as, with strictly textual patterns, there was a tendency for longer, more specific patterns. Patterns that were three tokens long were discarded, as they were too general. Also, the fact that the patterns at higher scoring levels (>0.5) on average were longer (six tokens long) than patterns with lower scoring suggests that specific patterns were better at finding opposites.

Acquired PoS pattern types did not differ from PoS pattern types found with adjective - adjective seeds and strictly textual pattern types, discussed in Chapter 4. Amongst the most productive pattern types were [*between* <*ANT*> *and* <*ANT*>], [*not* <*ANT*> *or* <*ANT*>], [*and* <*ANT*> *and* <*ANT*>].

In comparison with strictly textual patterns, the same set of 18 noun - noun seeds generated almost 20k less textual patterns. Those patterns extracted a much smaller number of pairs, namely, 2,019 pairs were found with textual patterns and 5,014 pairs were found with PoS patterns. Given that the results for PoS patterns had better precision, we conclude that surface PoS patterns provide a better means for finding opposites expressed by nouns.

Scoring	Pairs found with		
	6 seeds	12 seeds	18 seeds
≥0.9	52% (191)	49.7% (250)	47% (271)
≥0.8<0.9	15.5% (57)	17.1% (86)	18.4% (106)
≥0.7<0.8	16.1% (59)	16.5% (83)	15.6% (90)
≥0.6<0.7	13.1% (48)	13.1% (66)	13.8% (80)
<0.6	3.3% (12)	3.6% (18)	5.2% (30)
Total	*367*	*503*	*577*

Table 5.17: Total number of pairs per scoring level found with six, 12 and 18 verb - verb seeds by means of PoS patterns in TwNC.

5.4.3 Results for verb - verb seed pairs

Using a full version of TwNC, a set of six seeds extracted 367 unique pairs found five or more times, a set of 12 seeds found 503 pairs found more than five times and a set of 18 seeds extracted 577 unique pairs found more than five times. The summary of how many pairs were found with each seed set by means of PoS patterns per score level is given in Table 5.17. Similar to the results with adjective and noun seeds, more pairs were found with a larger set of verb - verb seeds and larger sets included all pairs found in the set of six seeds. Six seeds found 136 pairs less than 12 seeds and 210 pairs less than a set of 18 seeds. The difference in the number of found pairs between the set of 12 and 18 seeds was not so large - the set of 18 seeds found 74 more pairs than 12 seeds. This suggests that once there are enough seeds, the results do not change substantially. A similar difference remains among pairs at the highest score level ($\geq$0.9), where the set of six seeds found 59 pairs less than a set of 12 seeds and 80 pairs less than a set of 18 seeds; and the set of 12 seeds found 21 pairs less than the set of 18 seeds. Since the pairs found by the set of 18 seeds include all pairs found by smaller seed sets, these pairs will be discussed in detail. Pairs with scoring $<$0.6 were discarded.

5.4.3.1 Found pairs

Under assumption that opposites co-occur with each other within a sentence significantly often, we use significant co-occurrence as the first step to separate non-opposites from the results. As is shown in Table 5.18, out of the total 547 found pairs with the score $\geq$0.6, 81.3% co-occurred significantly more often than would be expected by chance (445 pairs). In comparison to significant co-occurrence of adjective - adjective and noun - noun pairs, verb - verb pairs co-occur significantly often less frequently. In

Scoring	Number of pairs	Significant co-occurrence
≥0.9	271	88.2% (239)
≥0.8<0.9	106	73.6% (78)
≥0.7<0.8	90	80% (72)
≥0.6<0.7	80	70% (56)
Total	*547*	81.3% (445)

Table 5.18: Number of unique pairs found by means of PoS patterns with 18 verb - verb seeds in the TwNC per scoring level, and number of pairs that co-occurred with each other sententially significantly more often than would be expected by chance.

particular, 92.5% of found adjective pairs and 94.4% of found noun pairs co-occurred with each other significantly often. This seems to indicate that a method based on surface patterns that are restricted to a certain number of linearly ordered tokens might be less applicable to finding opposites expressed by verbs than by adjectives and nouns.

More pairs had significant co-occurrence at higher score levels (88.2% or 239 pairs with the score ≥0.9), suggesting that there were more good candidates among them. Among discarded pairs were many verb - verb combinations with the verb *worden* "to become" that is often used to form the passive constructions, for example, *worden - wassen* "to become - to wash" as in *werd gewassen* "was washed". Thus, significant co-occurrence was helpful in eliminating non-opposites.

While we only consider significantly co-occurring pairs as candidate opposites, later in this section, we also examine the affect of significant co-occurrence on the performance of the algorithm by comparing precision scores for only those found pairs that co-occurred significantly often in the TwNC with precision scores for all found pairs (see Table 5.21 for further details).

All significantly co-occurring pairs were first evaluated by means of the available lexical resources for Dutch. The results are presented in Table 5.19. First, we examined how many of found pairs had both words present in CORNETTO and how many of them were linked as opposites (column 3).

Out of 239 pairs, 231 had both words present in CORNETTO (96.6%) and 27 of them were linked as opposites (11.7%). All but one of them were found among pairs with the highest score. In comparison, 160 adjective - adjective pairs (13.8% out of 1,155) and 36 noun - noun pairs (1.9% out of 1,851) were opposites according to this computational lexical resource.

Scoring	Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
≥0.98	137	19.4% (26/134)	16.8% (23)	24.8% (34)
≥0.96<0.98	30	0 (0/29)	3.4% (1)	3.4% (1)
≥0.94<0.96	36	3% (1/33)	2.8% (1)	5.6% (2)
≥0.90<0.94	36	0 (0/35)	5.6% (2)	5.6% (2)
Total	*239*	11.7% (27/231)	11.3% (27)	16.3% (39)

Table 5.19: Pairs with scoring ≥0.9 found with 18 verb - verb seeds in TwNC by means of PoS patterns and the number of pairs that were found in one or both of the lexical resources (CORNETTO and *Mijnwoordenboek.nl* (*MWB*)).

Also 27 pairs were opposites according to *MWB* but only 55.5% of them were the same opposites as the ones found in CORNETTO. As a result, the total number of opposites identified by one of the two or by both resources was 39 pairs, or 16.3% out of the total 239 pairs. Thus, using more resources for evaluation of automatically found candidate opposites is more productive than relying on one resource.

Among opposites identified only by CORNETTO, there were symmetric opposites *doorgaan - stoppen* "to go on - to stop", *gelijkspelen - verliezen* "to break even - to lose", *gelijkspelen - winnen* "to break even - to win" (24 in total) and asymmetric opposites (three in total). Among opposites identified only in *MWB* were pairs *praten - zwijgen* "to talk - to be still", *verhogen - verlagen* "to increase - to decrease". Opposites found in both resources included pairs *loslaten - vasthouden* "to release - to hold", *duwen - trekken* "to push - to pull". These examples demonstrate that there are no underlying theoretical differences among pairs found in one but not the other resource, suggesting that these resources supplement each other.

Next, all pairs were evaluated by three participants who were asked to classify each pair as an opposite or a non-opposite. Participants achieved a Fleiss's kappa score of 0.638, indicating substantial agreement. The agreement between participants in evaluation of verb - verb pairs was lower than that of adjective - adjective pairs (Fleiss's kappa score 0.74) but higher than that of noun - noun pairs (Fleiss's kappa score of 0.617). This shows that evaluation of noun - noun pairs was the most difficult task for the participants.

Out of 239 pairs, 36.4% (87 pairs) were judged as opposites by the majority vote, 61% of which received unanimous votes. The other 63.6% of pairs were judged as non-opposites, 83.5% of which received unanimous votes.

Among 53 unanimously judged opposites were mutual incompatibles like *bewon-*

Scoring level	Opposites		Non-opposites		Total
	by majority	unanimously	by majority	unanimously	
≥0.98	71.3% (62)	66.1% (41)	49.3% (75)	80% (60)	137
≥0.96<0.98	10.3% (9)	44.4% (4)	13.8% (21)	85.7% (18)	30
≥0.94<0.96	8.1% (7)	28.6% (2)	19% (29)	89.6% (26)	36
≥0.90<0.94	10.3% (9)	66.7% (6)	17.8% (27)	85.2% (23)	36
Total	*36.4% (87)*		*63.6% (152)*		*239*

Table 5.20: Percentage of pairs with scoring ≥0.9 extracted with 18 verb - verb seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

deren - haten "to admire - to hate", *rijden - stilstaan* "to drive - to stand still"; directional opposites *aankomen - vertrekken* "to arrive - to depart", *komen - gaan* "to come - to go" and other opposites like *verliezen - veroveren* "to lose - to conquer". Among 34 opposites that did not receive unanimous votes were pairs that are not mutually exclusive, for example, *leren - werken* "to study - to work" (one can study and work), *hopen - vrezen* "to hope - to fear"; pairs that have more than one counterparts, for example, *gelijkspelen - verliezen* "to break even - to lose" (the third opposite is *to win*), *blijven - gaan* "to stay - to go" (*to come*), *leven - sterfen* "to live - to die" (*to be born*). The pair *doorgaan - stoppen* "to go on - to stop", which was judged as opposites by the majority vote, seems to lack reversibility, by stopping one terminates the action rather than reverses it or does the opposite. The opposites *leiden - volgen* "to lead - to follow", and *besteden - sparen* "to spend - to save" also did not receive unanimous votes suggesting that classification of verb - verb pairs was not an easy task for the participants.

Recall that a total of 43 pairs found with verb - verb seeds by means of strictly textual patterns were judged as opposites by the majority vote and only 28 of them were expressed by verbs. Those 28 verb - verb pairs were also among pairs found by means of surface PoS patterns, which means that PoS patterns identified 62 more opposites expressed by verbs than did textual patterns. Among such pairs were opposites *slapen - waken* "to sleep - to wake up", *doorgaan - ophouden* "to continue - to stop", *haten - bewonderen* "to hate - to admire" and others.

Among 127 unanimously judged non-opposites many were frequently co-occurring words, for example, *stellen - vragen* "to raise - to question" (parsing error, it is actually a verb - noun pair *to ask questions* identified in patterns like [*time to ask questions*]), *grappen - grollen* "to joke - to gag", *puffen - persen*, "to wheeze - to push (to squeeze)"

(one has to do one or the other but not both), *slikken - stikken* “to swallow - to stitch” as in ‘*Het is slikken of stikken*’ (*you have to accept it*), *vergeten - vergeven* “to forget - to forgive” and others. There were also 17 pairs that contained the verb *worden* “to become”, used together with another verb in the passive voice. Some pairs belonged to contrastive sets like *horen - lezen* “to hear - to read” and *schrijven - praten* “to write - to speak”. Similar pairs were also among non-opposites that did not receive unanimous votes (25 pairs or 16.4% of all non-opposites), for example, pairs *horen - zien* “to hear - to see”, *luisteren - kijken* “to listen - to watch”.

To understand why these pairs were found by PoS patterns as candidate opposites and received scores $\geq$0.9, it is useful to look at some of the sentences in which these pairs were found.

For example, the pair *luisteren - kijken* “to listen - to watch”, which was judged as non-opposites by the majority vote, was found in the sentence:

- **luisteren - kijken**: Dat is een luisteraar die we hebben gevraagt die avond naar een programma op tv of radio te *kijken* of *luisteren*.

- *to listen - to watch*: That is the listener who we asked to *watch* a program on TV or to *listen* to a program on the radio tonight.

In this example, *to listen* and *to watch* are diametrically opposed as they refer to two opposite means of media branches, which in this context exclude one another. This is not recognized by the judges outside of the context, as in its more general sense, this pair also refers to various (more than two) ways of information, watching, reading, listening, seeing and so on. As a result, it was classified as non-opposites.

The pair *luisteren - praten* “to listen - to talk” was judged as opposites by the majority vote. Among sentences, in which this pair was found, was the following sentence:

- **luisteren - praten**: Marcos heeeft niet veel vertrouwen in de nieuwe president, Vicente Fox, “de man die veel *praat* maar weinig *luistert*”: we moeten veel lawaai maken, anders hoort hij ons niet.

- *to listen - to talk*: Marcos does not have much trust in the new president, Vicente Fox, “the man who *talks* a lot but does not *listen* much”: we should make a lot of noise, otherwise he cannot hear us.

Because the action of listening and talking is intuitively perceived as a bidirectional process, this pair was recognised as antonymous by the majority of participants. However, the way, in which the pairs *to listen - to talk* and *to listen - to watch* are used in the contrastive patterns, suggests that such pairs are similarly antonymous.

The examples above illustrate how difficult it can be for judges to recognize opposites in the absence of context. It also shows that theoretical approaches to antonymy should take antonym context-dependency into account. Another example, which can illustrate this point, is the pair *genezen - voorkomen* "to recover - to prevent". There is no diametric opposition between these two verbs and it is difficult to think of the context in which they are opposed, one can *prevent falling ill* so that there is no need *to recover*, but the pair *to fall ill - to recover* seems to be better, that is more direct, opposites in this case. As is shown in the following sentence, however, this pair is antonymous when it comes to the possible ways of handling medical symptoms. In this case, the prevention of a disease and the recovery from a decease are opposed with each other.

- **voorkomen - genezen**: Chinese arts hebben bij kuifapen een medicijn beproeven dat de longziekte sars zowel kan voorkomen als genezen.
- **to prevent - to recover**: Chinese doctors have tested a medicine that can prevent as well as treat the lung disease in macaques.

The pair was mostly found in the variations of the pattern [*zowel kan <ANT> als <ANT>*] "can <ANT> as well as <ANT>", which also found such opposites as *verliezen - winnen* "to lose - to win", *stijgen - dalen* "to increase - to decrease", *zingen - rappen* "to sing - to rap" and others.

Recall that we used CORNETTO to evaluate found pairs. Now that we have manual classification of the results, we can use it to evaluate this computational lexical resource by examining how many of verb - verb pairs judged as opposites by the majority vote had both words present in CORNETTO and how many of them were linked as opposites. Out of 87 pairs judged as opposites by the participants, 70 pairs (80.4%) were present in CORNETTO and 16 of them (22.8%) were linked as opposites. This means that the other 54 opposites are present in this resource but are not linked by the antonymy relation. Among missing opposites were pairs *vasthouden - loslaten* "to hold - to release", *besteden - sparen* "to spend - to save", *kopen - verkopen* "to buy - to sell" and *vergroten - verkleinen* "to increase - to decrease".

Scoring level	All found of pairs	Precision	Pairs with significant co-oc.	Precision
≥0.98	147	0.45	137	0.46
≥0.96 <0.98	37	0.22	30	0.18
≥0.94 <0.96	43	0.11	36	0.13
≥0.90 <0.94	44	0.25	36	0.23

Table 5.21: Precision scores based on the classification by three participants for pairs with scoring ≥0.9 which were overall found in TwNC (col. 2, 3) and only those that co-occurred with each other significantly often (col. 4, 5). Results found with 18 verb - verb seeds.

Top-k found pairs	Precision scores for textual patterns	Precision scores for PoS patterns
50	0.74	0.68
100	-	0.56
150	-	0.45
200	-	0.38

Table 5.22: Top-k pairs with scoring ≥0.9 extracted with 18 verb - verb seeds. Precision scores are based on the classification of pairs by three participants.

Significant co-occurrence was also used as the first step for evaluation of the results, by discarding non-significant pairs from the results. Now, by comparing precision scores for all found pairs with precision scores for pairs that had significant co-occurrence, we can evaluate how helpful is significant co-occurrence as a preliminary step for filtering out noise from the results. The scores are presented in Table 5.21.

As can be seen in the Table, significant co-occurrence has no positive effect on the precision.

Next, we examined precision scores for each top-k found pairs. The precision score for strictly textual patterns was available only for the top-50 pairs, as textual patterns found a total of 71 pairs with the score ≥0.9. Again the precision score for the top-50 pairs found by means of strictly textual patterns (0.74) is higher than the precision score for the top-50 pairs found by means of surface PoS patterns (0.68). The reason for that is because only 28 pairs found by means of textual patterns were expressed by verbs; the rest of the top-50 pairs were expressed by adjectives and nouns and comprised the most frequently found opposites, which were also present among the top found pairs in the results for all three seed sets. Thus, although the precision score even for the

top-pairs found by means of surface PoS patterns are lower than the precision score for the top-50 pairs found by means of strictly textual patterns, the results are better in that they only include verb - verb opposites.

Table 5.23 presents the top-50 pairs found by means of surface PoS patterns with 18 verb - verb seeds. As can be seen from the Table, almost all of the pairs were expressed by verbs. Two pairs were erroneously identified as verbs due to parsing errors. Namely, in the pair *stellen - vragen* "to ask - questions", the plural form of the noun *question* coincided with the infinitive form of the verb *to ask*. For the same reasons, the algorithm found the noun - noun pair *pieken - dalen* "peaks - valleys", which is often used as a fixed expression to mean *ups and downs*.

Among found pairs were also opposites from the original seed set, but not all of such pairs were unanimously recognized by the participants as opposites. In particular, the pairs *to end - to begin* and *to find - to lose* were judged as opposites by the majority vote. The pair *to ask - to answer* was classified as non-opposites by the majority vote. These inconsistencies in judgement show that also well-established opposites are not necessarily recognized by the participants in the classification tasks, when many pairs are presented and no context is provided. Thus, even canonical opposites need contextual support to be recognized as such in simplified classification tasks.

While such pairs as *to speak - to be still*, *to stay - to go away*, *to go - to come* were unanimously recognized as opposites, the pairs *to break - to make*, *to write - to read* and others were opposites by the majority vote and the pairs *to hear - to see*, *to read - to hear* and *to write - to say* were judged by the majority vote as non-opposites. Again this suggests that there seems to be a continuum with mutually exclusive opposites on its one side and opposites that have more than one counterpart on the other.

In comparison to other seed sets, verb - verb seeds found the smallest number of pairs. Table 5.24 gives an overview of precision scores for the top-k pairs for the sets of six, 12 and 18 seeds. As can be seen, the best results were achieved with the largest seed set, suggesting that opposites expressed by verbs are less frequent and therefore more seeds tend to find more patterns and consequently pairs.

5.4.3.2 Acquired verb - verb PoS patterns

A total of 10,848 unique patterns with part-of-speech information were acquired with the set of 18 verb - verb seeds. In comparison, the set of adjectival seeds generated 18,983 patterns and the set of nominal seeds generated 65,867 patterns. The fact that

Dutch	English	Judged as opposites
toenemen - dalen	to increase - to decrease*	yes unanimously
afstoten - aantrekken	to reject - to recruit	yes unanimously
eindigen - beginnen	to end - to begin*	yes by majority vote
sluiten - opennen	to close - to open*	yes unanimously
vinden - verliezen	to find - to lose*	yes by majority vote
eten - drinken	to eat - to drink	yes by majority vote
schrijven - lezen	to write - to read	yes by majority vote
exporteren - importeren	to export - to import*	yes unanimously
aanvallen - verdedigen	to attack - to defend*	yes unanimously
stijgen - dalen	to increase - to decrease*	yes by majority vote
winnen - verliezen	to win - to lose*	yes unanimously
horen - zien	to hear - to see	no by majority vote
beantwoorden - vragen	to answer - to ask*	no by majority vote
opstaan - vallen	to rise - to fall	yes by majority vote
huilen - lachen	to cry - to laugh*	yes unanimously
kopen - verkopen	to buy - to sell*	yes unanimously
ontkennen - bevestigen	to deny - to confirm*	yes unanimously
bieden - loven	to offer - to praise	no unanimously
geven - nemen	to give - to take*	yes unanimously
mislukken - slagen	to fail - to succeed*	yes unanimously
laten - doen	to let - to do	yes unanimously
stellen - vragen	to pose - questions	no unanimously
vallen - staan	to fall - to stand	no unanimously
doen - zeggen	to do - to say	no unanimously
wonen - werken	to live - to work	no unanimously
passen - meten	to fit - to measure	no unanimously
huren - kopen	to rent - to buy	yes unanimously
doen - denken	to do - to think	yes by majority vote
wikken - wegen	*part of col. expression* to consider	no unanimously
gaan - komen	to go - to come	yes unanimously
blijven - weggaan	to stay - to go away	yes unanimously
waken - slapen	to wake up - to sleep	yes unanimously
buigen - barsten	*part of col. expression* bend or break	no by majority vote
stikken - slikken	*part of col. expression* to accept	no by majority vote
breken - maken	to break - to make	yes by majority vote
lezen - horen	to read - to hear	no unanimously
trekken - duwen	to pull - to push	yes unanimously
uitsluiten - bevestigen	to exclude - to confirm	no unanimously
drinken - roken	to drink - to smoke	no unanimously
zwijgen - spreken	to be still - to speak	yes unanimously
staan - zitten	to stand - to sit	yes by majority vote
kiezen - delen	to choose - to share/to divide	no unanimously
pieken - dalen	peaks - valleys	yes unanimously
afnemen - toenemen	to decline - to increase	yes unanimously
schrijven - zeggen	to write - to say	no by majority vote
begraven - cremeren	to bury - to cremate	yes by majority vote
landen - opstijgen	to land - to take off	yes by majority vote
lopen - fietsen	to walk - to cycle	no unanimously
beleggen - sparen	to invest - to save	yes by majority vote
liggen - zitten	to lay - to sit	no by majority vote

Table 5.23: Fifty top pairs found with 18 verb - verb seeds by means of PoS patterns and their classification according to three judges.

Top-k found pairs	Pairs found with 6 seeds	Pairs found with 12 seeds	Pairs found with 18 seeds
50	0.6	**0.68**	**0.68**
100	0.42	0.51	**0.56**
150	0.34	0.41	**0.45**
200	-	0.35	**0.38**

Table 5.24: Top-k pairs with scoring $\geq$0.9 extracted with six, 12 and 18 verb - verb seeds. Precision scores are based on the classification of pairs by three participants and lexical resources.

seeds expressed by verbs generated the least number of patterns, twice as few as the same set of seeds with strictly textual patterns suggests that verb - verb pairs are less likely to co-occur together in surface patterns, which are specific enough to contain both verbs and at the same time general enough to contain a range of different verb - verb pairs.

The shortest patterns were four tokens long (14.2%), and the longest patterns were seven tokens long (20%). Thus, there was a tendency for longer, more specific patterns. Patterns that were three tokens long were discarded as they were too general. Also, the fact that the patterns at higher scoring levels (>0.5) on average were longer (six tokens long) than patterns with lower scoring suggests that specific patterns were better at finding opposites.

Recall that in Chapter 4 we discussed that the most productive pattern types among strictly textual patterns identified by adjective - adjective and noun - noun seeds differed from the most productive pattern types identified by verb - verb seeds. Namely, the most productive pattern type among adjectives and nouns was [*between* <*ANT*> *and* <*ANT*>] and for verbs [*(n)either* <*ANT*> *(n)or* <*ANT*>] and [*to* <*ANT*> *or to* <*ANT*>]. We argued that these differences were found due to the differences in the main discourse functions (Jones [2002]) of opposites expressed by different syntactic categories. The types of surface PoS patterns identified by means of adjective and noun seeds were similar to the types of strictly textual patterns identified by the same seed sets. Analysis of pattern types of surface PoS patterns identified by means of verb - verb seeds showed that similar to previous results, the most productive PoS pattern type was [*(n)either* <*ANT*> *(n)or* <*ANT*>]. Also variations of the generic pattern [<*ANT*> *and* <*ANT*>] were frequently identified amongst the most productive patterns. While these differences have not been fully addressed in the previous studies on discourse functions of opposites, our findings show that the syntactic category of the opposites might play

a role in the most predominant discourse functions they indicate.

5.5 Discussion

In this chapter we presented a pattern-based method for finding opposites expressed by a specific syntactic category. This method was based on the algorithm presented in Chapter 4 but instead of identifying strictly textual patterns, the algorithm automatically generated PoS patterns that contain part-of-speech information of the target word pairs using a small set of seeds. We expected that PoS patterns would be able to find fewer candidate pairs than strictly textual patterns but that more of those pairs would be valid opposites.

Our results show that strictly PoS patterns successfully find opposites expressed by all three part-of-speech categories, showing that a pattern-based method that uses automatically found patterns offers a promising means for finding opposites in the future work. PoS patterns particularly outperform strictly textual patterns in finding opposites expressed by nouns and verbs, suggesting that when the most frequent adjective - adjective opposites do not influence the results, less frequently but equally good opposites expressed by nouns and verbs can be found in the top results.

Contrary to our assumption that PoS patterns would find fewer candidate pairs than strictly textual patterns, the results show that PoS patterns find many more pairs than textual patterns. Moreover, more of them are opposites according to the computational lexical resources as well as manual evaluation.

5.5.1 Automatically found opposites

As expected, the PoS patterns successfully dealt with both shortcomings of the strictly textual patterns. First of all, they did not find any cross-categorical pairs, with the exception of a few parsing errors like the pair *stellen* - *vragen* “to pose - questions”. Second, as has been mentioned above, they found a much larger number of opposites for each syntactic category. For example, using the set of 18 verb - verb seeds, strictly textual patterns found a total of 71 pairs whereas PoS patterns found 239 pairs. Out of the 71 pairs, 43 pairs were judged as opposites by the majority vote but only 28 of them were expressed by verbs. On the contrary, out of the 239 pairs found with PoS patterns, 87 pairs were judged as opposites, and they were all expressed by verbs. Thus, adding

syntactic information to a pattern-based method is very beneficial for finding opposites expressed by verbs.

PoS patterns were also successful at finding adjective - adjective and noun - noun pairs. Namely, 517 pairs found with 18 adjective - adjective seeds and 399 pairs found with noun - noun seeds were judged as opposites by the majority vote. In comparison, with strictly textual patterns, 208 pairs found with 18 adjective - adjective and 220 pairs found with 18 noun - noun seeds were opposites according to the majority vote and many of these pairs were overlapping across results for different seed sets (for example, 33% of top-250 pairs found with adjectival and nominal seed sets were the same, see Table 4.30 for further details). Again, this shows that PoS patterns outperform textual patterns at finding opposites expressed by syntactic categories other than adjectives.

Found pairs expressed by adjectives achieved the highest inter-annotator agreement among three seed sets. In particular, the Fleiss's kappa score for classification of 1,563 pairs found by adjective seeds was 0.74, the Fleiss's kappa score for classification of 2,132 pairs found by noun seeds was 0.617 and the Fleiss's kappa score for classification of 239 pairs found by verb seeds was 0.638. In other words, it is much easier for the participants to classify opposites and non-opposites for pairs expressed by adjectives than by verbs and nouns.

Recall that in a similar classification of pairs found with 18 adjective - adjective seeds with strictly textual patterns, participants achieved a Fleiss's kappa score of 0.66 (see Section 4.4.1.1, Chapter 4 for details). This is an interesting result, given that in the latter case, the participants had to evaluate 475 pairs, whereas in the former case they classified 1,563 pairs. That is, they demonstrated a much higher degree of agreement, even though they had to evaluate three times as many pairs. The main difference between the results in the latter and former studies is that this time all 1,563 found pairs were expressed by adjectives, whereas strictly textual patterns found pairs expressed not only by adjectives but also by nouns and sometimes verbs. Again, this shows that participants find it easier to decide whether a pair is antonymous or not for strictly adjectival pairs.

Noun - noun opposites vs co-hyponyms. Participants had most difficulties classifying noun - noun pairs. This is not surprising, given that many of such pairs would traditionally be considered as co-hyponyms rather than standard opposites. However, such pairs also differ from standard co-hyponyms in that they are naturally contrasted.

The members of this subclass seem to be the traditional subclass of multiple incompatibles (Lyons [1977]). Because these particular pairs seem to function in the

newspaper corpus like other antonymous pairs, our results provide evidence for arguing that such mutual incompatibles are a subtype of opposites.

It is crucial to note that not all co-hyponyms are naturally contrasted and that this property may not always be the main import of the pair. Within the contrasting contexts created by the patterns, certain co-hyponym pairs may be more likely to be seen as opposites rather than as sisters of the same hypernym. As a result, our judges were often split as to whether a given pair should be classified as opposites or non-opposites, especially given that all found pairs were presented to them outside of the context. For example, pairs *church - state*, *fact - fiction*, *family - friend*, *citizen - authorities*, *luck - wisdom*; see Table 1.15 for other examples.

It is unclear if it is the pattern type or specific context that has a contrastive function that then emphasizes incompatible features of certain word pairs that are otherwise co-hyponyms, or if it is the case that the same word pair is ambiguous for both an antonym and a co-hyponym function. The former explanation seems more likely because the patterns found seem to be very effective in emphasizing contrasts.

Not all co-hyponyms function naturally in contrastive contexts this way. For example, *table*, *seat*, *bureau* and *dresser* are all co-hyponyms of furniture according to WordNet 2.0 (Fellbaum [1995]), yet we would consider none of these pairs contrastive in the same way that a pair like *country-side - city* is. But our method does not extract such pairs because, unlike opposites, they do not co-occur with each other significantly more often than chance would predict. The co-hyponyms we found seem to allow contrasts suggesting that it is useful to treat them as one lexical class of word pairs with opposites.

The question of whether or not it is the pattern or inherent characteristics of the pair that is the root of incompatible meaning is related to the question of pairs' context-dependency. Murphy [2003] has suggested that it is the patterns that pairs occur in that are responsible for their contrastive meaning, and the work of Jones (2002) has focused on contrastive functions based on patterns, and not on opposites themselves. For the class of co-hyponyms, identifying the context they are used in seems to be essential for their interpretation.

Canonicity. Another area in which the role of patterns has been argued to play a role is antonym canonicity. Recall that previous studies have suggested that the differences between canonical and non-canonical opposites are reflected in terms of their breadth of co-occurrence, or the range of patterns in which a given pair co-occurs. Our results show that non-typical opposites like *white - red* and *conservative - progressive* were

found with the most productive PoS patterns that also found well-established canonical opposites like *rich - poor*. This seems to indicate that significant co-occurrence in patterns of incompatibility is a property of non-canonical pairs as much as it is of canonical opposites. We mentioned earlier that Jones et al. [2007] examined the range of patterns, and not pattern types, only in relation to canonical opposites, neglecting non-canonical opposites. However, it seems that neither significant co-occurrence nor the breadth of co-occurrence are sufficient for assessment of antonym canonicity.

5.5.2 Automatically found PoS patterns and their types

Our results show that a pattern-based method can be applied to finding opposites expressed by all three part-of-speech categories. Our automatically identified patterns are more specific than manually-selected patterns. As a result, automatically identified patterns do not find as much noise as strictly textual patterns in the top of results.

The unexpected result in relation to pattern identification is that, contrary to our expectations, PoS patterns found more opposites for each syntactic category than surface textual patterns, even though overall the same seed sets identified more textual patterns than part-of-speech patterns. For example, 18 adjective ? adjective seeds found approximately 30k textual patterns and almost 19k part-of-speech patterns. In other words, they identified 10k more textual patterns than PoS patterns. Yet, textual patterns found a total of 1,049 candidate pairs, where as PoS patterns found 3,275 candidate pairs, which is three times as much. Moreover, three times as many pairs found with PoS patterns (134 pairs in total) received the highest score of one, compared to only 40 pairs found with textual patterns.

A further look at the top pairs showed that on average, pairs extracted with part-of-speech patterns were found in more patterns than pairs extracted with textual patterns. For example, the pair *zwak ? sterk* "weak - strong" was found with 71 textual patterns and 332 part-of-speech patterns. What this seems to suggest is that initially the set of 18 adjective ? adjective seeds identified more textual patterns as mentioned above. However, as such patterns do not restrict the syntactic category of the target words, the seeds find many more patterns, including those that contain noun ? noun and verb ? verb pairs. When these patterns are automatically scored, **many** of them get lower scores. For example, 5,154 PoS patterns were discarded because they had an automatic scoring below the threshold of 0.1. In comparison, 13,042 textual patterns were discarded because of the automatic scoring below 0.1. Also, there were twice as many

PoS patterns with automatic scoring $\geq$0.6 then textual patterns. This consequently led to the higher scoring of pairs.

Whereas we were able to identify opposites expressed by adjectives and nouns, we were not very successful with verbs. It can be that surface patterns identified by our algorithm were too specific to contain a wide range of verb - verb opposites.

Also the differences in pattern types might play a role. Recall, that the most productive pattern types identified by verb - verb seeds were not the same as the ones identified by adjective - adjective and noun - noun seeds, which implies that the main discourse functions of verb - verb opposites might be different from those of adjectival and nominal opposites. If this is the case, it can be that opposites expressed by verbs are less likely to co-occur together within short proximity than opposites expressed by adjectives and nouns. If this is the case, a pattern-based method based on the linear ordering of words will be unable to identified many good instances of verbal antonymy. This shortcoming of surface patterns can be overcome in a pattern-based method that uses dependency patterns, that capture syntactic dependencies rather than surface proximity among words. This is the method we explore in the next chapter.

CHAPTER 6

Performance of dependency patterns for finding opposites

We present an automatic method for extraction of opposites by means of *dependency patterns*, that is patterns that contain syntactic relations between words.[1] Using several sets of seeds that express one of the three part-of-speech categories, we identify the best dependency patterns and use them to find novel opposites.

Similar to the surface part-of-speech patterns, dependency patterns find opposites expressed by a target syntactic category defined by the seed sets, which eliminates noise caused by cross-categorical pairs. But unlike surface patterns, which rely on the linear ordering of words in close proximity, dependency patterns can deal with the so-called long-distance dependencies, that is, cases when opposites co-occur in a sentence too far from each other and, as a result, they cannot be found by means of strictly textual patterns nor by part-of-speech patterns. Also, given that recent studies by Snow et al. [2005], Snow et al. [2008] show that dependency patterns outperform other meth-

[1] The material presented in this chapter has been published as Anna Lobanova, Gosse Bouma and Erik Tjong Kim Sang (2010) Using a Treebank for Finding Opposites. In: Eds. Markus Dickinson, Kaili Müürisep and Marco Passarotti. Proceedings of TLT9, Tartu, Estonia, pp.139-150.

ods at finding meronyms and hyponyms, it is important to study the performance of dependency patterns at finding opposites.

The results presented in this chapter show that dependency patterns find novel opposites. The system performed best with adjectival seeds, followed by nouns and verbs. However, the most frequently found dependency patterns are too general and extract a lot of noise. We conclude that while syntactic information helps to identify opposites expressed by nouns and verbs, there is no overall improvement with dependency patterns over surface part-of-speech patterns.

In an ongoing debate on the usefulness of syntactic information in relation extraction (Snow et al. [2005], Snow et al. [2006], Tjong Kim Sang and Hofmann [2009], and others), our findings support earlier claims of Tjong Kim Sang and Hofmann [2007] who argue that syntactic information does not improve the performance of pattern-based methods for hyponym-hypernym extraction. Taking computational costs of full syntactic parsing into consideration, we conclude that part-of-speech patterns discussed in Chapter 5 are more attractive for antonym harvesting.

6.1 Inspirations for the present study

This is the first study that uses dependency patterns for finding opposites. As will be discussed in Section 6.1.1, one of the main incentives to try out dependency patterns for finding opposites is their flexibility in dealing with opposites that cannot be found by means of surface patterns because they are located too far away from each other. Further, as will be discussed in Section 6.1.2, previous studies on dependency patterns in relation extraction suggest that such patterns outperform surface patterns in finding pairs of hypernyms - hyponyms and meronyms. Interestingly, as will be discussed below, not everyone agrees with the latter claim, arguing that surface patterns with part-of-speech information are as good as dependency patterns. And because they do not require as much syntactic preprocessing as dependency patterns, patterns with part-of-speech information can be applied to larger corpora. Studying the performance of dependency patterns on finding opposites is, therefore, also useful for answering a general question about the differences in performance of different pattern types in relation extraction.

6.1.1 Dependency patterns versus surface patterns

6.1.1.1 Limitations of surface patterns for finding opposites

In Chapters 4 and 5 we presented two pattern-based methods that use *surface* patterns for finding opposites. One of the main properties of all kinds of surface patterns is that they rely on the linear ordering of words in a sentence. They also contain no syntactic information, except the part-of-speech categories in the case of part-of-speech patterns.

As has been discussed in Chapter 2, the idea of using patterns for studying opposites can be traced back to the work of Justeson and Katz [1991], who were the first to suggest that significantly co-occurring opposites '... are usually paired, and in these cases they are *commonly* [my emphasis] found in conjoined phrases that are identical or nearly identical, word for word, except for the substitution of one antonym for the other' [p. 10]. Together with significant co-occurrence, phrasal repetition, or so-called "surface syntactic similarity", was used to explain antonym canonicity. Namely, the high association of opposites was viewed as a result of their significant sentential co-occurrence with each other in surface patterns, in which they could be substituted for one another. The substitution was an alignment mechanism so that opposites could be interchanged with one another in an otherwise identical or near-identical context, that is, a pattern. This caused a strong association between them.

Later, such patterns, or antonym "near-identical contexts" were studied by Jones [2002], who analysed 3000 newspaper sentences with well-established opposites. He shifted the focus from antonym canonicity to the functions of opposites in discourse, studying different types of surface patterns in which opposites co-occurred. Interestingly, one of his main findings was that in almost 40% of the identified sentences, opposites did *not* co-occur in any identifiable *surface* patterns. Instead, opposites often co-occurred in parallel constructions to emphasize an opposition between words or phrases that would not be contrasted otherwise, as in the example below:.

(1) In Russia, the "oligarchs" typically acquired wealth by trading in commodities *purchased* domestically at *regulated* prices, and then *sold* abroad at *deregulated* prices.

Note that in most of such cases, it was not possible to identify a productive *surface* pattern because the opposites were found linearly too far away from each other. Recall that strictly textual patterns (Sections 4.4.1.2 and 4.4.2.2) and surface part-of-speech

patterns (Sections 5.4.1.2 and 5.4.2.2) were on average six tokens long. This is because longer patterns are too specific to be found more than once, containing different pairs of opposites. Thus, the 'linear nature' of surface patterns poses a potential limitation on the recall of the surface pattern-based algorithm.

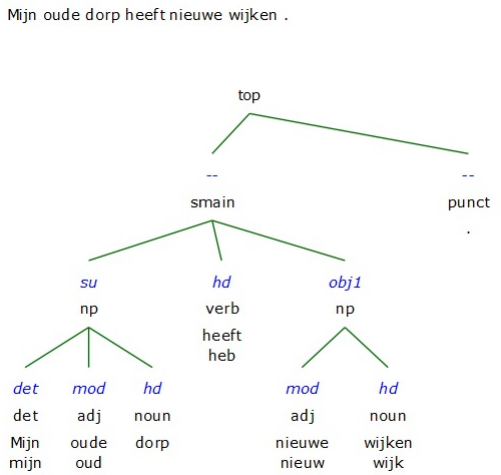
It is important to point out that we refer to this limitation of surface patterns as potential. This is because it is yet necessary to establish whether opposites found far away from each other are qualitatively different from opposites found in surface patterns. Recall that Jones [2002] studied types of patterns in regard to the types of discourse functions of opposites rather than the types of opposites found in different pattern types. More importantly, remember that Jones found both identifiable surface pattern types as well as parallel constructions with long-dependencies using the same set of opposites. This seems to indicate that surface patterns might be sufficient for finding different types of opposites. In other words, it can be that the same opposites found far away from each other are also found in close proximity. In this case scenario, dependency patterns might be more useful for identification of a wider range of discourse functions of opposites in the text. We explore this by comparing opposites found by means of dependency patterns with opposites found by means of surface PoS patterns, examining differences and similarities between them.

6.1.1.2 Dependency patterns: Introduction

Dependency patterns are acquired from treebank data, and contain syntactic relations between elements of a sentence. As a result, dependency patterns are less specific and less constrained by the surface order of elements than surface patterns. The dependency pattern *Verb1:conj* ← *of* → *conj:Verb2* (Verb1:conj ← or → conj:Verb2), for example, links the two verbs in (1), representing the shortest path between them in the dependency tree. It is an equivalent of a surface PoS pattern [*<ANT1/Verb> or <ANT2/Verb>*].

In order to generate dependency patterns, sentences have to be parsed. A *dependency tree* for sentence (2) below, produced by the Alpino parser for Dutch (van Noord [2006]), is shown in Figure (2):

(2) Mijn oude dorp heeft nieuwe wijken. (NRC, Dec 20, 2000)
My old village has new areas.



By convention, syntactic relations in such trees can be presented as tuples of the form ($word_1$, CAT1: *Relation*: CAT2, $word_2$). Each tuple contains the following information:
$word_1$ - the lemmatized form of the head of the relation;
$word_2$ - the lemmatized form of the dependent;
CAT1 and CAT2 - part-of-speech categories of $word_1$ and $word_2$;
Relation - a dependency relation between $word_1$ and $word_2$.

For example, the dependency relations between words in example (2) include

(*have*, VERB: *Subject*: NOUN, *village*)
(*have*, VERB: *Direct Object*: NOUN, *area*)
(*village*, NOUN: *Modifier*: ADJ, *old*)
(*area*, NOUN: *Modifier*: ADJ, *new*)
(*my*, DET: *Determiner*: NOUN, *village*)

A dependency pattern is then defined as a set of partially underspecified dependency relations.

For example, the following pattern can be constructed from the aforementioned tuples for finding the pair *new* - *old*:

[<ANT_1/Adj>:Mod←*village*:Sub←*have*→DirObj:*area*→Mod:<ANT_2/Adj>].

6.1.1.3 The benefits of syntactic information in relation extraction

Since performance of automatic sentence parsing systems is steadily improving (Surdeanu et al. [2008]), it is becoming increasingly plausible to use full parsing in relation extraction (Berland and Charniak [1999], Snow et al. [2005], van der Plas and Bouma [2005], Tjong Kim Sang and Hofmann [2007] among others).

The extent to which syntactic information is beneficial, however, is still an open question. For example, Tjong Kim Sang and Hofmann [2009] compared two automatic methods for hypernym-hyponym extraction for Dutch (for example, pairs *meubels - stoel* "furniture - chair", *vervoermiddel - auto* "vehicle - car"). In one method they used dependency patterns, while the other method relied on surface PoS patterns. They found that PoS patterns performed as well as dependency patterns given a 20% larger corpus. The main differences in the recall were due to the inability of the surface PoS patterns to deal with long-distance dependencies and due to the parsing errors introduced in the dependency patterns. Since the part-of-speech tagging does not require a lot of preprocessing, the authors argued that PoS patterns offer a better method for finding hyponyms across vast data collections.

Results in an earlier study of Snow et al. [2005], however, showed that dependency patterns outperformed surface PoS patterns for hypernym-hyponym extraction in English. However, their PoS-pattern-based classifiers used a much smaller set of features, that is, PoS patterns, than their classifiers with dependency patterns. As a result, the comparison was not well-balanced and it is difficult to draw any conclusions.

6.1.1.4 The benefits of syntactic information for finding opposites

In the previous chapter, we have shown that surface PoS patterns can be used to successfully find opposites in a newspaper corpus. Using dependency patterns for the same task on the same corpus will shed more light on the effectiveness of syntactic information in extraction of opposites. But comparing the performance of surface-based methods with a dependency-pattern method for antonym extraction has also another substantial advantage, namely, opposites provide an opportunity to study pattern-based methods for automatic extraction of a relation expressed by different part-of-speech categories.

An important difference between antonymy as opposed to hyponymy (for example, *rose - flower*) and meronymy (for example, *finger - hand*) is that only the former relation holds between words that can be expressed not only by nouns (for example,

beginning - end) but also by adjectives (for example, *happy - sad*), verbs (for example, *give - take*) and other categories (for example, adverbs like *quickly - slowly* and prepositions *up - down*). Because of this, examining a dependency-based method for antonym harvesting is useful not only for understanding the benefits of syntactic information for relation extraction, but also for identifying the extent to which dependency patterns differ from surface patterns at finding pairs of opposites that belong to different part-of-speech categories. It can be that dependency patterns are more productive than surface patterns at extraction of opposites expressed by verbs. In this case, previous studies did not find this difference as they were interested in finding only noun - noun pairs.

It can also be that we will not find any differences in the performance of a dependency pattern-based algorithm with the seed sets of different syntactic categories. For example, Tjong Kim Sang and Hofmann [2007] compared a PoS pattern [*N such as N*] with its dependency pattern equivalent and found that pairs that were missed by this PoS pattern and found by the equivalent dependency pattern, were found by other PoS patterns. Thus, given a wide range of PoS patterns used in their study, pairs missed by one pattern seemed to be found by other types of PoS patterns.

6.1.2 Dependency patterns in relation extraction

6.1.2.1 Original work on dependency patterns

Dependency patterns were originally used in Question Answering (QA) systems and later in automatic extraction of lexico-semantic relations, particularly hyponymy and meronymy.

In QA, dependency paths have been successfully used to extract, for example, information like [*X writes Y*] is the same as [*X is the author of Y*] (Lin and Pantel [2001]). Jijkoun et al. [2004] and Bouma et al. [2005] show that using syntactic information in QA systems for Dutch improves the results.

In relation extraction, Snow et al. [2005] proposed to use generic dependency patterns for finding hyponyms. This was the first study that showed that classifiers trained on dependency patterns outperformed classifiers that used manually crafted surface patterns with part-of-speech information originally proposed in Hearst [1992]. In their approach, noun-noun pairs were collected from a fully parsed six million corpus of newspaper texts. Using WordNet (Fellbaum [1998]), all unique pairs were classified as either known hypernym or known non-hypernym pairs. Given sentences in which

these pairs occurred, dependency patterns were generated and patterns that occurred with at least five unique pairs were used to construct feature count vectors for each noun pair that occurred with at least five unique patterns. A number of classifiers were trained on these features and evaluated using a 10-fold cross validation on a set of 5,387 manually annotated noun pairs. The performance of their best logistic regression classifier achieved an F-score of 0.27. It outperformed a classifier based on manually crafted patterns from Hearst [1992] that obtained the lowest F-score of 0.14. The authors argued that their results indicate that dependency patterns based on syntactic information are not only useful for identifying hyponymy relations but they are better at hypernym-hyponym extraction than methods that use manually-selected surface PoS patterns.

Following [Snow et al. [2005]]'s approach, McNamee et al. [2008] used dependency patterns to find named entity hyponyms like *Jamaica - island*, *Hilton-hotel - hotel*, which are often not covered by lexical resources but can be very useful for QA systems. They reported a 9% improvement in performance of a QA system that uses extracted pairs.

Dependency patterns have also been used for finding meronyms. In particular, Ittoo and Bouma [2010] used a minimally-supervised bootstrapping algorithm to find meronyms of different types, for example, a *member-of* type as in *player - team*, or a *structural part-of* type as in *engine - car*. Their method was based on the *Espresso* algorithm originally proposed by Pantel and Pennacchiotti [2006] for finding meronyms and hyponyms in English, but whereas Pantel and Pennacchiotti [2006] generated surface PoS patterns, Ittoo and Bouma [2010] generated dependency patterns. Recall that an *Espresso*-like algorithm automatically identifies generic (i.e., high recall and low precision) patterns automatically. Starting with a small set of seeds, all sentences in which seed pairs co-occur are extracted and used to generalize patterns. All patterns are automatically evaluated using an association measure between a given pattern and highly reliable instances based on pointwise mutual information Church and Hanks [1990]. The top-10 best patterns are used to find new pairs. Extracted pairs are also evaluated using an association score between a given pair and a highly reliable pattern. Ittoo and Bouma [2010] used the algorithm on a 450 million words newspaper corpus of Dutch that we use in this study. Based on the evaluation by two judges, they achieved a 60% - 80% precision on top 500 extracted meronym pairs.

In the study presented in this chapter we followed [Ittoo and Bouma [2010]]'s approach and developed a similar *Espresso*-like algorithm. Due to the constraints on

the available computational resources, it was not possible to conduct a study using a methodology analogous to the algorithms presented in Chapters 4 and 5. Moreover, using an *Espresso*-based algorithm has several additional advantages.

First, an *Espresso*-based algorithm can be applied to a vast amount of data without loss of computational power and time. Second, a principled measure of pattern and instance reliability enables it to have both high precision and high recall. Third, similar to the pattern-based methods presented in earlier chapters, this algorithm requires no human annotation. Finally, as Pantel and Pennacchiotti [2006] mention themselves, it must be applicable to a wide variety of relationships. Previous studies have shown that this algorithm outperforms other methods for finding a number of lexical-semantic relations, including meronymy and hyponymy in English and Dutch. It is, therefore, necessary to examine its performance for finding opposites and to compare the results with previous findings.

Note that some research has questioned whether or not syntactic information adds anything to the pattern-based methods for relation extraction. In particular, Tjong Kim Sang and Hofmann [2009] replicated Snow et al. [2005]'s approach on Dutch and compared it with a method that uses surface PoS patterns. The experiments on hyponym-hypernym extraction were conducted using two Dutch corpora: a collection of newspaper texts from the Twente News Corpus (approximately 26 million sentences) and a collection of Dutch Wikipedia texts (approximately five million sentences). No significant differences were found between the two types of extraction patterns. The largest effect was found for Wikipedia texts, where dependency patterns found 23% more related pairs than surface PoS patterns. The authors argued that this effect could be overcome for surface PoS patterns by adding 43% extra data.

The examples of the best PoS and dependency patterns found by Tjong Kim Sang and Hofmann [2009]'s algorithm illustrate that there are no differences in the types of generated patterns, suggesting that differences in the results were not due to the genre of the corpus. Overall, they found no significant differences between the two approaches and concluded that using more data rather than more syntactic information can substantially improve the results. While pre-processing of surface PoS patterns requires tokenization, part-of-speech tagging and lemmatization, dependency patterns need an additional costly step of complete dependency parsing that introduces additional errors. Since shallow parsing can easily be applied to extensive data collection, the authors conclude that approaches based on lexico-syntactic patterns are as useful as dependency patterns at considerably cheaper processing costs.

Dependency patterns have not yet been used for finding opposites. The main goal of the experiments presented in this chapter is to fill in this gap and to examine whether syntactic information improves the performance of a pattern-based method for finding opposites. Our second goal is to examine whether dependency patterns preform equally well for finding opposites expressed by all three part-of-speech categories. By using dependency patterns with seeds that belong to different part-of-speech categories we can investigate whether syntactic information is more useful for pairs *and* relations that belong to a particular part-of-speech category.

6.2 Assumptions

Based on the previous findings discussed above, we have the following assumptions for the results of the algorithm that uses dependency patterns for finding opposites:

1. **Automatic identification of opposites**:
    - opposites found automatically will be expressed by all three part-of-speech categories;
    - cross-categorical pairs will not be found by the algorithm.
2. **Automatic identification of dependency patterns**:
    - given a large enough corpus, it is possible to identify useful dependency patterns automatically;
    - automatically generated dependency patterns can successfully find good opposites;
    - dependency patterns will find a similar range of opposites in comparison with strictly textual patterns and surface PoS patterns.

6.3 Method

6.3.1 Corpus

We used a 450 million words version of the Twente Nieuws Corpus of Dutch (TwNC, Ordelman [2002]) that consisted of 26 million sentences. The corpus consists of news-

wire texts from five daily Dutch newspapers.[1] The corpus was syntactically parsed by Alpino, a parsing system for Dutch aimed at parsing unrestricted texts (van Noord [2006]). The accuracy of Alpino is over 90% (tested on a set of 2,256 newspaper sentences (van Noord [2006]), which is comparable to the state-of-the-art parsers for English (Collins [1996], Charniak [2000], Lin and Pantel [2001]).

6.3.2 Seeds

The same seed sets that were used in the experiments with strictly textual (Chapter 3) and surface PoS (Chapter 4) patterns were used in this work. They are summarized in Table 4.1 in Chapter 4.

6.3.3 The Algorithm

Our method is based on the well-known minimally-supervised bootstrapping algorithm *Espresso* (Pantel and Pennacchiotti [2006]). First, using seed pairs as tuples, dependency patterns that contained both words of a pair were extracted from the treebank. Patterns that were found only once or twice were discarded. Next, found dependency patterns were automatically scored. The reliability of a pattern p, $r_\pi(p)$, given a set of input pairs I was calculated as its average strength of association across each input (seed) pair i in I, weighted by the reliability of each input pair i, $r_\iota(i)$:

$$r_\pi(p) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{max_{pmi}} * r_\iota(i) \right)}{|I|}$$

where $pmi(i,p)$ is the pointwise mutual information score Church and Hanks [1990] between a pattern and an input pair, and $\max_{pmi}$ is the maximum pointwise mutual information score between all patterns and all pairs. The reliability of initializing seed pairs was set to 1. Next, the top-k most reliable patterns were used to find new candidate pairs. We set the number of the initial set of top patterns to 10, adding one extra pattern at each iteration. Because we use a much larger corpus than Pantel and Pennacchiotti [2006], we do not retrieve additional instances of patterns from the Web. We also do not use a discounting factor suggested in Pantel and Ravichandran [2004] and used

[1] Namely, *Algemeen Dagblad*, *NRC Handelsblad*, *Parool*, *Trouw* and *Volkskrant*.

in Pantel and Pennacchiotti [2006] to control for the bias of *pmi* towards infrequent events. Instead we remove patterns and pairs that occur once or twice.

The reliability of found pairs, $r_\iota(i)$ was estimated as follows:

$$r_\iota(i) = \frac{\sum_{p \in P} \left(\frac{pmi(i,p)}{max_{pmi}} * r_\pi(p) \right)}{|P|}$$

where P is the set of top-k found patterns.

The top-100 found pairs were used as new seeds in the next iteration. The process was repeated iteratively until at least 500 new pairs were acquired.

6.3.4 Evaluation

The results of the current dependency-based algorithm comprise a ranked list of at least 500 best pairs for each seed set. These lists were evaluated in the same manner as findings presented in Chapters 4 and 5.

First, we examined how many of found pairs co-occurred significantly often. Next, the results were evaluated against CORNETTO, a computational lexical resource for Dutch, and against the online dictionary *Mijnwoordenboek.nl*.

Finally, all found instances were classified by three judges in a 'Yes/No' classification task. Participants were presented with one pair at a time and were asked to classify each pair as opposites or non-opposites. We report a Fleiss's kappa score for inter-annotator's agreement (Randolph). A score between 0.61 and 0.8 is considered to indicate a substantial agreement. Using manual classification and CORNETTO, we estimated precision scores.

6.4 Results

The results are presented for each part-of-speech category separately. First, we discuss the results for adjective - adjective seeds (Section 6.4.1), then we discuss the results for noun - noun seeds (Section 6.4.2) and finally we present the results for verb - verb seeds (Section 6.4.3).

Set1 & Set2	Only Set 1	in Set 2	Overlap	Total number of found pairs
6 & 12 seeds	494	500	13	1007
6 & 18 seeds	498	510	9	1017
12 & 18 seeds	76	82	437	596

Table 6.1: Number of pairs found with adjective - adjective seed sets of different sizes and the overlap between the sets.

6.4.1 Results for adjective - adjective seed pairs

Adjective-adjective seed sets of different sizes yielded very different results. The overview of how many pairs were found with each seed set and how many of them overlapped, that is, found by more than one seed set, is presented in Table 6.1. As can be seen, the overlap between the results found with six and 18 seeds was particularly small. Namely, out of 507 unique pairs found with the set of six seeds, only 13 pairs (3%) were among 513 pairs found with 12 seeds and only nine pairs (2%) were among 519 pairs found with the set of 18 seeds. Even the top-pairs found with six seeds were very different from the top-pairs found with 18 seeds. In comparison, the overlap between the results found with 12 and 18 seeds consisted of 437 pairs. This demonstrates that the number of seeds drastically affects the results and using larger seed sets for identification of dependency patterns leads to more consistent results.

Among pairs that were found by all three seed sets were pairs *dicht - open* “closed - open”, *duur - goedkoop* “expensive - cheap”, *dik - dun* “thick - thin”, *langzaam - snel* “fast - slow” and a few others. The results with larger seed sets contained typical as well as non-typical opposites like *normaal - verhoogd* “normal - enhanced”, *onzichtbaar - zichtbaar* “visible - invisible”. A manual analysis of pairs found by means of six seeds showed that except for a small set of well-known opposites, six seeds found mostly noise.

This seems to suggest that unlike with surface PoS patterns discussed in Chapter 5, a small set of six seeds is not sufficient for finding reliable dependency patterns. Recall that all pairs found by six seeds with PoS patterns were also found by the sets of 12 and 18 seeds (see Section 5.4.1 for more details).

A further look at the types of dependency patterns acquired by seed sets of different sizes revealed that there was a striking difference in the types of dependency patterns generated by means of six seeds as opposed to dependency patterns generated by the sets of 12 and 18 seeds. In particular, dependency patterns found by means of six

Number of found pairs	Pairs with significant co-occurrence
519	88.4% (459)

Table 6.2: Number of pairs found by means of dependency patterns with 18 adjective - adjective seeds and the percentage of pairs that co-occurred in the TwNC significantly more often than would be expected by chance.

seeds were more specific and contained information over more syntactic relations than dependency patterns found by means of 12 and 18 seeds. Thus, with dependency patterns, more seeds lead to a similar recall but higher precision.

Recall, that with strictly textual patterns and with surface PoS patterns, larger seed sets also increased the recall. For example, using PoS patterns, the set of 18 adjective-adjective seeds found five times as many pairs with scoring $\geq$0.6 and three times as many pairs with scoring $\geq$0.9 as the set of six seeds. However, although more pairs were found, the precision score for the results found with different seed sets was higher only for the top-k pairs.

Because of the large overlap in the results between 12 and 18 seeds, we will discuss only pairs found with the largest set of 18 seeds in detail.

6.4.1.1 Pairs found with 18 adjective - adjective seeds

In relation to significant co-occurrence, our results showed that dependency patterns found fewer significantly co-occurring adjective - adjective pairs than surface PoS patterns. Recall that one of the prerequisites for two words to be antonymous, is for them to co-occur with each other within a sentence significantly more often than would be expected by chance Charles and Miller [1989]. As can be seen in Table 6.2, 88.4% of pairs (459 pairs) found by means of the 18 adjective - adjective seeds co-occurred sententially in the newspaper corpus significantly more often than would be allowed by chance. In comparison to the results found with surface patterns, this number is lower. In particular, using the same seed set, 98.3% (or 475 pairs) of pairs with the score $\geq$0.9 found with strictly textual patterns and 95.4% (or 1,563 pairs) of pairs with the score $\geq$0.9 found with surface PoS patterns co-occurred with each other within a sentence in the same newspaper corpus significantly often. Because pairs found with textual patterns contained all three part-of-speech categories, these results are not comparable. But the fact that surface PoS patterns found three times as many adjective -

adjective pairs as dependency patterns yet contained a larger percentage of pairs with significant co-occurrence (92.5%) than dependency patterns (88.4%) suggests that dependency patterns are more likely to find less frequently co-occurring pairs. Among such pairs were *moeilijk - saai* "difficult - boring", *dramatisch - strafbaar* "dramatic - punishable", *doeltreffend - vlug* "efficient - quick" and other non-opposites. As will be discussed later in Section 6.5, such pairs were found by means of very general dependency patterns.

Among adjective - adjective pairs that co-occurred significantly often were opposites *direct - indirect* "direct - indirect", *abstract - figuratief* "abstract - figurative", *gezond - ziek* "healthy - ill", *nationaal - internationaal* "national - international", *militair - politiek* "military - political", *tijdelijk - permanent* "temporary - permanent" as well as non-opposites *handig - verstandig* "handy - wise", *geestig - ontroerend* "witty - touching", *werkloos - ziek* "jobless - ill", *eerlijk - vrij* "honest - free". Again, this illustrates that significant co-occurrence is a good preliminary cue for separating opposites from other frequently co-occurring pairs but that it is not a sufficient cue for eliminating all non-opposites from the results.

Out of 459 significantly co-occurring pairs found with 18 adjective - adjective seeds, 86.3% (396 pairs) had both words present in CORNETTO, and 18.2% of them (72 pairs) were linked as opposites in this computational lexical resource (see Table 6.3 for details). Twenty-one opposites (29.2%) were linked with each other asymmetrically. For example, while *mannelijk* "male" was among opposites of *vrouwelijk* "female", *vrouwelijk* "female" was not among opposites for *mannelijk* "male". Similarly, *optimistisch* "optimistic" was among opposites of *pessimistisch* "pessimistic" but *pessimistisch* "pessimistic" was not among opposites of *optimistisch* "optimistic". Other asymmetric pairs included opposites *correct - incorrect* "correct - incorrect", *horizontaal - verticaal* "horizontal - vertical", *dom - slim* "stupid - clever" and so on. Among symmetric pairs were opposites *klassiek - modern* "classical - modern", *triest - vrolijk* "sad - happy", *licht - zwaar* "light - heavy" and others. As has already been said in Chapter 4, these examples show that the asymmetry does not reflect any underlying theoretical assumptions and it is rather a manifestation of the inconsistency in the encoding of opposites in CORNETTO.

The next step was to evaluate found pairs by examining how many of them were opposites according to one or both lexical resources.

Overall, 100 pairs of 459 extracted pairs, or 19.3%, were opposites according to one or both resources.

Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
459	18.2% (72/396)	17% (78)	21.8% (100)

Table 6.3: Percentage of found pairs listed as opposites in CORNETTO (col. 2), in *Mijnwoordenboek.nl* (col. 3) or in both resources (col. 4). The second number in column 2 represents the total number of found pairs for which both words are present in CORNETTO.

Among pairs that were present in CORNETTO but not linked as opposites were co-hyponyms like *dom - eigenwijs* "foolish - stubborn", *belangrijk - interessant* "important - interesting". There were also opposites, for example, pairs *lelijk - mooi* "ugly - beautiful", *schuldig - onschuldig* "guilty - innocent", *modern - ouderwets* "modern - old-fashioned", *veilig - onveilig* "safe - unsafe", *echt - nep* "real - fake" and others. This again illustrates how beneficial automatic extraction of opposites is for the enrichment of CORNETTO. Later on, to get a better understanding of the extent to which antonymy is represented in CORNETTO, we examine how many of pairs judged as opposites by the majority vote are found in CORNETTO and marked as opposites. In other words, instead of using CORNETTO for evaluation of our results, we will use our results, classified by judges, for evaluation of the coverage of antonymy in CORNETTO.

The online dictionary *Mijnwoordenboek.nl* (*MWB*) contained slightly more pairs identified as opposites: 17% or 78 pairs. Sixty-four percent of them (50 pairs) were also identified as opposites in CORNETTO; these were the overlapping pairs. Among pairs listed as opposites in both resources were pairs *gehaat - geliefd* "hated - beloved", *dik - dun* "thick - thin", *expliciet - impliciet* "explicit - implicit", *links - rechts* "left - right" and others. Pairs listed as opposites only in *MWB* contained pairs *dood - levend* "dead - alive", *gezond - ziek* "healthy - sick", *heet - koud* "hot - cold". Among opposites found only in CORNETTO were pairs *abstract - figuratief* "abstract - figurative", *gedwongen - vrijwillig* "compulsory - voluntarily". There were no principled differences between the pairs classified as opposites only in CORNETTO or only in *MWB*, suggesting that the inconsistencies are due to the overall under-coverage of opposites in the resources.

In comparison with the results found by means of part-of-speech patterns, out of 1,563 adjective - adjective pairs with the score $\geq$0.9, almost 74% had both words present in CORNETTO with almost 14% of them, namely, 160 pairs, listed as opposites. Another 49 pairs were found only in *MWB*, comprising a total of 172 adjective - adjective pairs marked as opposites by one or both of the resources. This shows that

Opposites		Non-opposites		Total
by majority	unanimously	by majority	unanimously	
36.8% (169)	82.8% (140)	63.2% (290)	82.8% (240)	459

Table 6.4: Percentage of significantly co-occurring pairs found with dependency patterns with 18 adjective - adjective seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

based on the lexical resources, part-of-speech patterns found more opposites.

It is then important to examine how many pairs found by means of dependency patterns were classified as opposites by judges, as it might be that dependency patterns found qualitatively different pairs from pairs extracted with part-of-speech patterns. In a 'Yes/No' classification task, we asked three participants to classify each pair as an opposite or a non-opposite. The results are summarized in Table 6.4. Recall that unanimous votes indicate how many pairs were judged as opposites or non-opposites by all three participants. These votes are included in the column with votes by the majority vote.

Out of the total 459 pairs, 36.8% (169 pairs) were judged by the majority vote as opposites, the other 63.2% (290 pairs) were judged by the majority vote as non-opposites. Among 140 pairs unanimously judged as opposites were pairs *waar - onwaar* "true - false", *aangeboren - aangeleerd* "innate - learnt", *automatisch - handmatig* "automatic - manual".

Among 240 unanimously judged non-opposites were pairs *onduidelijk - onjuist* "unclear - unfair", *tactisch - technisch* "strategic - technological" as well as pairs *laag - middelhoog* "low - medium-high", *juridisch - politiek* "judicial - political". Analysis of the sentences in which these pairs were found, showed that pairs like *unclear - unfair* and *strategic - technological* were mostly found in general dependency patterns equivalent to surface patterns [<*ANT/Adj*> *and* <*ANT/Adj*>] and [<*ANT/Adj*> *or* <*ANT/Adj*>]. This illustrates that among productive dependency patterns identified by the algorithm were generic patterns that also tend to find noise.

Although the pairs *low - medium-high* and *judicial - political* were not recognised as opposites by any of the participants, it is possible to find contexts in which these two examples are contrasted. The pair *low - medium-high* was not recognised as contrastive most likely because the two words do not define the two opposite points on the scale of

height, a property exhibited by well-established pairs of opposites like *high - low*, *tall - short*, and other gradable adjectival pairs.

Also the pair *rood - blauw* "red - blue" was unanimously discarded by the participants as non-opposites. However, the pair *red - blue* is contrastive in many different contexts. For example, it was contrastive in the context of the colour of blood, in particular, it is common to associate the red colour with oxygenated blood and the blue colour with deoxygenated blood. The pair was also contrastive in the contexts of warm and cold colours, high and low temperatures, and even in the reference to political systems, identifying capitalism with *blue* and socialism with *red* as in *Immers, de nazaten van de rode socialisten en de blauwe kapitalisten vloeiden samen tot een paars kabinet! 'At the end, the descendants of the red socialists and the blue capitalists merged into a purple cabinet!'*. In the same paragraph of the newspaper article in which this sentence was found, the author discusses how a seemingly impossible to overcome ideological gap between the two political parties had been closed in the previous decade. In other words, there was an assumption that the two movements are incompatible.

Such associations are not captured in the association tests that deal with the most common opposites like *rich - poor*, *tall - short*, and others. The fact that these pairs were also not recognized as opposites by our judges points out that these opposites are context-dependent and without the context it is difficult to rely on intuition for their recognition. Also, given that such pairs co-occur significantly often and that they are found in patterns that also find well-established opposites, it seems that the significant co-occurrence and the substitutability are insufficient for separating well-established opposites, recognized as such outside of any context, and context-dependent opposites, not recognized as such based on the judges' intuition alone.

While such pairs are ignored in theoretical classifications of opposites, they are as useful for NLP applications, for example, for finding the discourse relation of Contrast, as readily recognized pairs. Unfortunately, this context-dependency makes the evaluation of the results found by our algorithm difficult and somewhat misleading, as many good pairs are treated as non-opposites and discarded from the results. To resolve this problem it is important to examine in the future in psycholinguistic experiments whether participants are more likely to recognize opposition when pairs are presented with some context and how much context would be sufficient in such tasks.

Looking at the pairs that did not receive unanimous votes either as opposites or as non-opposites, it appears that these two categories are similar. Among 29 opposites that did not receive unanimous votes were pairs *akoestisch - elektrisch* "acoustic - electric",

bestaand - nieuwgebouwd “existing - newly built”, *blind - doof* “blind - deaf”. Most of such pairs were not categorized as opposites by all three judges because they belong to a category with more than two members. As a result, they were probably treated as co-hyponyms. The context in which these pairs were found, however, illustrates their contrastiveness. For example, the pair *acoustic - electric* is contrastive, for example, in relation to the types of guitars as in the sentence ‘*Which is better to learn to play on ... an electric guitar or an acoustic guitar?*’. The pair *existing - newly built* is contrastive, for example, in relation to the tax payment differences when purchasing a newly built or an existing house.

Also the pair *leeg - vol* “empty - full” did not receive unanimous votes as an opposite, suggesting that participants did not always recognize well-established opposites. Although they could come back to any pair and change their answer as many times as needed, participants almost never used this option.

There were also contrastive pairs among 50 non-opposites by the majority vote. For example, this group contained pairs that refer to a scale but not its opposite poles, for example, pairs *anderhalf - half* “one and a half - half”, *neutraal - positief* “neutral - positive”. This group also contained pairs that, similarly to co-hyponyms, belonged to a category with multiple members. In such cases, they were mutually incompatible, that is, one could not be both at the same time. For example, the pair *civiel - militair* “civil - military”, or *medisch - psychisch* “medical - psychological”. Again, this suggests that the contrast indicated by these two words is not readily recognized by the judges without the necessary context.

Overall, although we used two mutually exclusive categories for the evaluation of the results, namely, opposites and non-opposites, the results of the judges seem to reflect the spectrum of different degrees of antonymy. Opposites that received unanimous votes represent the most readily recognized opposites (for example, *white - black*). These pairs were followed by the opposites that did not receive unanimous votes. Such pairs were opposites but not as typical, or as canonical, as unanimously judged pairs (for example, *existing - newly built*). Next followed the pairs that were judged as non-opposites by the majority vote. They consisted of contrastive pairs that belong to a category with more than two members, but could still be used as opposites, given the context (for example, the pair *informative - entertaining*). Without the context, they were often treated as co-hyponyms. However, it is important to keep in mind that our algorithm did not find co-hyponyms like *chair - table*, which are very unlikely to be found in contrastive contexts. This illustrates that co-hyponyms found in dependency

patterns that were identified by means of antonymous seeds find inherently contrastive pairs.

Finally, the pairs unanimously judged as non-opposites represent the actual noise found by the algorithm. Often, these are frequently co-occurring pairs, for example, *strategic - technological*. However, also good opposites were present in this category, suggesting that sometimes participants failed to recognize opposites without the context.

Despite aforementioned limitations, manual classification is still the most reliable means of evaluation of the results. In our case, participants recognized 140 unanimously judged opposites, or 169 opposites by the majority vote, whereas both lexical resources, taken together, identified only 100 opposites among found pairs.

We can use the results of manual classification to evaluate the coverage of opposites in CORNETTO. In particular, we can examine how many of opposites by the majority vote are present in this resource, and how many of them are opposites. Out of the 169 opposites that received the majority vote, 90% (152 pairs) had both words present in CORNETTO and 44.1% of them (67 pairs) were linked as opposites. This means that more than half of valid opposites extracted by the adjective - adjective seed pairs and recognized by the participants as antonymous are missing the label 'opposites' in the most recent computational lexical resource for Dutch.

Recall that the program was iterated over five times and at each iteration, top-100 found pairs were added to the seed set used in the previous round and an additional pattern was added to the top-k best patterns set to ten in the first iteration (see Section 6.3.3). We will now examine how these settings influenced the results. Did more seeds give higher precision? Did more patterns lead to higher recall? The results are summarized in Table 6.5.

The table presents the number of significantly co-occurring pairs (column 2) found per iteration (column 1), the percentage of how many of them were judged as opposites by the majority vote (column 3), and how many were opposites according to CORNETTO (column 4), based on the number of pairs for which both words are present in the lexical resource. The last two columns represent precision scores based on the manual evaluation alone (in which case only unanimously judged pairs were taken into consideration) or based on the combination of manual evaluation and CORNETTO, so that unanimously judged opposites and/or opposites found in CORNETTO were treated as true positives and unanimously judged non-opposites were treated as false positives.

As can be seen, the best precision score was achieved for the 113 pairs found in the

Iteration	Found pairs	Opposites by majority vote	Opposites in CORNETTO	Precision scores based on judges	Precision scores based on judges & CORNETTO
1	113	63.7% (72)	35.9% (37/103)	0.68	**0.69** (69)
2	199	53.3% (106)	27.7% (49/177)	0.54	**0.55** (98)
3	291	46.7% (136)	24.4% (64/262)	0.48	**0.51** (126)
4	376	40.9% (154)	20.9% (69/329)	0.41	**0.43** (139)
5	459	36.8% (169)	18.2% (72/396)	0.37	**0.39** (152)

Table 6.5: Number of pairs with significant co-occurrence found per iteration, the percentage of how many were opposites according to the majority vote, the percentage of how many were present in CORNETTO and linked as opposites, and precision scores (based on unanimous votes and unanimous votes combined with opposites according to CORNETTO). The results are presented for the pairs found by means of 18 adjective - adjective seeds.

first iteration. The precision score varied between 0.68 and 0.69, showing that treating pairs marked as opposites in CORNETTO as unanimously judged opposites helps to better assess the performance of the algorithm. Overall, the precision scores based on manual evaluation alone were lower than those based on manual evaluation and the resource. For example, in the third iteration, the precision score based on manual evaluation, namely 0.48, was more than 15% lower than the precision score based on the combination of the two. One of the reasons for this is that more seeds are able to find less typical opposites which are not unanimously recognized by the participants. Relying on an additional resource, in which such pairs are partially listed as opposites, can help to reduce the negative effect of the limitations of manual classification on the evaluation of the results.

Recall that for the results of surface patterns we calculated the precision score based on the top-k found pairs. Although all pairs found in one iteration were consequently found in the following iteration, the ordering of the pairs changed. Table 6.6 presents precision scores for the first 150, 200 and 250 pairs found at each iteration, as well as the precision score for the same top-k pairs found by means of part-of-speech patterns, described in Chapter 5.

Two important findings can be drawn from the results presented in Table 6.6. The first is that with each consequent iteration, as the number of seeds increased, the number of unanimously judged opposites in the top-k number of pairs increased as well, leading to a higher precision score. For example, the first 100 pairs found in iteration five contained ten more opposites, unanimously judged by the participants or marked as such in the CORNETTO, than the first 100 pairs found in iteration two. The second

Top-k found pairs	Precision scores for pairs found with dependency patterns in iteration 2	iteration 3	iteration 4	iteration 5	PoS patterns -
150	0.58 (78)	0.61 (81)	0.64 (87)	**0.65** (88)	0.57 (78)
200	-	0.57 (100)	0.58 (103)	**0.59** (104)	0.53 (98)
250	-	0.53 (116)	0.53 (115)	**0.55** (119)	0.5 (112)

Table 6.6: Precision scores for top-k pairs found in different iterations by means of dependency patterns (col. 2 - 5) and found by means of part-of-speech patterns (col. 6). The number of the unanimously judged opposites in each set is given in brackets. All pairs were found by means of the same set of 18 adjective - adjective seeds.

finding is that dependency patterns outperformed part-of-speech patterns, discussed in Chapter 5. In each set of top-k found pairs, dependency patterns identified more unanimously judged opposites and/or opposites present in CORNETTO than did part-of-speech patterns.

Interestingly, the overlap between top-k opposites found by means of dependency patterns and PoS patterns was small. For example, out of the top-200 pairs found with part-of-speech patterns and the first 200 pairs found in the fifth iteration with dependency patterns, 67 pairs were the same. This means that 66.5% of pairs in each set were different pairs.

This seems to suggest that different kinds of patterns tend to find a different range of opposites. However, a closer examination revealed that out of the 175 opposites (judged unanimously or by the majority vote and/or marked as opposites in CORNETTO) found by means of dependency patterns in one of the five iterations, 138 pairs, or 79%, were also found among pairs with the scoring $\geq$0.9 identified by means of part-of-speech patterns. Among 37 opposites found only by means of dependency patterns many were erroneously parsed as adjectives, for example, *tegen - voor* "against - in favour", *ontvangen - verzenden* "to receive - to send", *rijden - stilstaan* "to ride - to stay still" and others. On the other hand, 379 opposites present among the results with the scoring $\geq$0.9 found with part-of-speech patterns were not found by means of dependency patterns.

This means that the main difference between the two pattern-based methods lies in the productivity of patterns. Namely, using the same corpus and the same seed set, dependency patterns extract fewer pairs but with more opposites among top results. Part-of-speech patterns extract many more candidate pairs in order to identify the same

opposites. One of the possible reasons for such variations might be due to the differences in methodology, since the method presented in this chapter allows limiting the number of found pairs, and at the same time it uses a greater number of seeds in consequent iterations. This ensures that the small number of the best identified dependency patterns is capable of finding a larger number of opposites in the top results. The exact types of the most productive dependency patterns will be discussed next.

6.4.1.2 Dependency patterns acquired with adjective - adjective seeds

It is interesting to study dependency patterns acquired with adjective - adjective seeds for several reasons. Recall that one of the reasons for using dependency patterns was their ability to deal with opposites that sententially co-occur far away from each other, so surface patterns that could capture them cannot be constructed. It is, therefore, interesting to see whether dependency patterns differ from surface patterns in their length and specificity. Secondly, it is also interesting to see whether dependency patterns added at later iterations were different from the ones acquired at the initial iterations, as precision for the top-k pairs increased with each consequent iteration. Finally, it is useful to examine the types of dependency patterns and to compare them with the types of surface part-of-speech patterns due to the difference between them described in the previous section.

The three top patterns with the highest scoring used throughout five iterations were 'equivalents' of the following surface PoS patterns: [<*ANT/Adj*> *or* <*ANT/Adj*>], [<*ANT/Adj*> *as well as* <*ANT/Adj*>] and [*neither* <*ANT/Adj*> *nor* <*ANT/Adj*>], followed by such patterns as [<*ANT/Adj*> *and* <*ANT/Adj*>], [<*ANT/Adj*> *versus* <*ANT/Adj*>], [*more* <*ANT/Adj*> *than* <*ANT/Adj*>], [<*ANT/Adj*> *and* <*ANT/Adj*>] and others. These dependency patterns were not exactly equivalent to the surface PoS patterns as they were more general and did not correspond to the surface ordering of the words in a sentence. Because of that, instead of focusing on their length, we can instead compare their specificity and types.

With respect to the specificity, as these examples show, found dependency patterns are very general. In particular, two very general patterns, referred to as *generic* in the original work of Pantel and Pennacchiotti [2006] for having high recall and low precision, were [<*ANT/Adj*>*:conj* $\leftarrow$ *of* $\rightarrow$ *conj:*<*ANT/Adj*>] equivalent to the surface pattern [<*ANT/Adj*> *or* <*ANT/Adj*>] and [<*ANT/Adj*>*:conj* $\leftarrow$ *and* $\rightarrow$ *conj:*<*ANT/Adj*>] equivalent [<*ANT/Adj*> *and* <*ANT/Adj*>]. These patterns were discarded by both al-

Set1 & Set2	Only in Set1	Only in Set2	Overlap	Total unique pairs
6 & 12 seeds	369	375	138	882
6 & 18 seeds	384	395	123	903
12 & 18 seeds	143	148	370	662

Table 6.7: Number of pairs found by means of dependency patterns acquired with noun - noun seed sets of different sizes and the overlap between them.

gorithms with surface patterns, which preferred more specific patterns. The fact that we find generic patterns among the best dependency patterns might also indicate the differences between the two methods used for the assessment of pattern reliability.

As to the types of patterns, most of the dependency patterns could be classified according to the types identified in Jones [2002]. For example, patterns [*<ANT/Adj> and / or / as well as <ANT/Adj>*] belong to the Coordinated type. Interestingly, the algorithm identified the pattern [*<ANT/Adj> versus <ANT/Adj>*], which is always contrastive but not very frequent but it did not identify the equivalent of one of the most productive surface patterns - [*between <ANT/Adj> and <ANT/Adj>*].

6.4.2 Results for noun - noun seed pairs

Similar to the results with adjective - adjective seeds, the size of the noun - noun seed set led to differences in the results. Again, the overlap between pairs found by means of the smallest seed set and pairs found with larger seed sets was small. As is summarized in Table 6.7, although the overlap between noun - noun seed sets was bigger than that between adjective - adjective sets, the least number of common pairs was found between the set of six seeds and the other two sets. In particular, out of 507 pairs found with six seeds, 138 pairs (27%) were also found with the set of 12 seeds and 123 pairs (24%) were also found with the set of 18 seeds. Again, very few of the pairs found with the set of six seeds were opposites. The largest overlap was found among pairs found with 12 and 18 seeds. In particular, 370 pairs (72%) found with 12 seeds were also found with 18 seeds. This shows that using larger seed sets with dependency patterns leads to more consistent results also for pairs expressed by nouns.

The analysis of the patterns acquired by means of the small set of six seeds and the larger sets of 12 and 18 seeds showed that there was almost no overlap between acquired patterns. Whereas six seeds identified more specific patterns, 12 and 18 seeds identified mostly generic patterns.

Number of found pairs	Pairs with significant co-occurrence
518	91.5% (474)

Table 6.8: Number of pairs found by means of dependency patterns with 18 noun - noun seeds and the percentage of pairs that co-occurred in the TwNC significantly more often than would be expected by chance.

In the rest of this section we discuss in detail the results found by means of 18 noun - noun seeds.

6.4.2.1 Pairs found with 18 noun - noun seeds

As the first step, we examined how many of found pairs co-occurred with each other within a sentence in the TwNC significantly more often than would be expected by chance. The results are presented in Table 6.8. Out of 518 unique pairs found by the end of the fifth iteration, 91.5% (474 pairs) co-occurred within a sentence significantly often. Thus, more pairs found with dependency patterns with the noun seed set had significant co-occurrence than pairs found with adjective seed set (88.4% out of 519 found pairs). It could be that dependency patterns find more noun - noun pairs that are likely to frequently co-occur with each other at longer linear distances than adjective - adjective pairs.

In comparison with the results found with surface PoS patterns, 94.4% of found noun - noun pairs, 4,805 pairs, co-occurred in the corpus significantly more often than would be expected by chance.

Among pairs with significant co-occurrence were opposites *criticus - sympathisant* "critic - sympathiser", *innerlijk - uiterlijk* "interior - exterior", *opdrachtgever - opdrachtnemer* "contractor - employee", *consument - producent* "consumer - producer", *groep - individu* "group - individual", *hart - hoofd* "heart - head" as well as non-opposites like *gewoonte - taal* "custom - language", *automobilist - huiseigenaar* "car owner - house owner", *uitkomst - verloop* "outcome - process". Among pairs without significant co-occurrence were *coach - vriend* "coach - friend", *asielzoeker - scholier* "refugee - pupil", *cola - kunst* "soft drink - art". These examples show that significant co-occurrence can be used to automatically filter out partial noise from the results.

Next, we examined how many of found pairs that co-occurred significantly often were opposites according to CORNETTO and *MijnWoordenboek.nl*. The results are summarized in Table 6.9.

Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
474	2.9% (11/372)	9.5% (45)	9.9% (47)

Table 6.9: Percentage of found pairs listed as opposites in CORNETTO (col. 2), in *Mijnwoordenboek.nl* (col. 3) or in both resources (col. 4). The second number in column 2 represents the total number of found pairs, which have both words present in CORNETTO.

Overall, 47 pairs, 9.9% of 474 found pairs with significant co-occurrence, were opposites according to one or both lexical resources. In comparison, 100 adjective - adjective pairs found with dependency patterns (or 19.3% of 459 pairs with significant co-occurrence) and 103 noun - noun pairs, or 4.8% of the total 2,132 pairs, found with PoS patterns were opposites according to these lexical resources. One of the reasons for this can be a lower representation of noun - noun opposites in lexical resources. We will examine this later on in this section when looking at how many of manually classified opposites are found in CORNETTO as opposites. Yet, the fact that 103 noun - noun pairs found with PoS patterns were opposites according to the same resources seems to suggest that the two types of patterns find different kinds of pairs. We will also look into this option by examining the overlap in the results found by means of the same set of 18 noun - noun seeds but different pattern types.

Out of 474 pairs found with 18 noun - noun seeds, 78.5% (372 pairs) had both words present in CORNETTO, of which 2.9% (11 pairs) were linked as opposites. This means that out of 474 pairs, only 2.1% (11 pairs) are confirmed opposites according to CORNETTO. Two pairs were linked as oppostionites asymmetrically, for example, *koper - verkoper* "buyer - seller". Among symmetric pairs were opposites *kou - warmte* "coldness - warmth", *allochtoon - autochtoon* "foreigner - indigenous", *invoer - uitvoer* "import - export", *binnenland - buitenland* "inland - abroad", *duif - havik* "dove - hawk" and others. Again, it is difficult to identify any underlying theoretical differences between symmetric and asymmetric opposites, thus, the asymmetry is a matter of inconsistent encoding of opposites in CORNETTO. Interestingly, the opposites like *dove - hawk* are non-conventional, difficult to classify by any existing categories of opposites, showing that the choice of opposites represented in CORNETTO is difficult to systematise.

In comparison to pairs found by means of adjective seeds, CORNETTO was not as helpful for identification of opposites found by means of noun seeds. In particular, 72

pairs (or 18.2% of 396 pairs present in CORNETTO) were opposites among pairs found with adjective seeds and only 11 pairs (or 2.9% of 372 pairs present in CORNETTO) were opposites among pairs found with noun seeds. This result is similar to the result found with surface PoS patterns, where out of 1,033 pairs found with 18 noun - noun seeds that had both words present in CORNETTO, only 2.8% (29 pairs) were linked as opposites. The fact that we find similar results also for noun - noun pairs found by means of dependency patterns highlights the gaps in the coverage of noun - noun opposites this computational lexical resource. As will be discussed further, automatic extraction of opposites can greatly help to fill in such gaps with reliable pairs of opposites.

Among pairs that were present in CORNETTO but not linked as opposites were *trendvolger - trendsetter* "trend follower - trendsetter", *inkomst - uitgave* "income - expense", *droom - werkelijkheid* "dream - reality", *huurwoning - koopwoning* "rental apartment - bought apartment". All these pairs indicate opposition, especially when used in comparative and contrastive contexts.

Mijnwoordenboek.nl had more instances of found pairs. Namely, 9.5% (45 pairs) were opposites according to *MWB*. This number includes nine pairs also identified as opposites in CORNETTO. Examples of opposites found only in *MWB* are *jongen - meisje* "boy - girl", *gevolg - oorzaak* "result - cause", *min - plus* "minus - plus", *mislukking - succes* "failure - success", *toekomst - verleden* "future - past". The fact that *MWB* also identified the majority of noun - noun opposites found with PoS patterns shows that using more than one lexical resource for evaluation of automatically found opposites is very beneficial.

Next, we asked three native speakers to classify found pairs as opposites or non-opposites. In this task, participants achieved a Fleiss's kappa score of 0.67, which indicates sufficient agreement between participants. Nevertheless, this score is lower than the agreement score achieved by the participants in the evaluation of adjective - adjective pairs (0.76). This indicates that it was more difficult for the participants to classify noun - noun pairs. This is also supported by a lower agreement score among participants for the evaluation of noun - noun pairs found be means of surface PoS patterns (with the Fleiss's kappa score of 0.617).

The results of the classification task are summarized in Table 6.10. Columns with unanimous votes indicate how many pairs were judged as belonging to one of the target categories by all three participants. Unanimous votes are included in the numbers for the majority vote.

Opposites		Non-opposites		Total
by majority	unanimously	by majority	unanimously	
29.7% (141)	73% (103)	70.3% (333)	81.1% (270)	474

Table 6.10: Percentage of significantly co-occurring pairs found with dependency patterns with 18 noun - noun seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

Out of the total 474 pairs with significant co-occurrence, almost 30% of pairs were judged by the majority vote as opposites, which means that the other 70% of pairs were judged by the majority vote as non-opposites. More than 70 percent of opposites received unanimous votes. Among 103 unanimously judged opposites were pairs *opluchting - teleurstelling* "relief - disappointment", *kind - volwassene* "child - adult", *tegenstander - vriend* "adversary - friend", *consument - producent* "consumer - producer" and others.

Among 270 unanimously judged non-opposites, 81.1% of all non-opposites, were frequently co-occurring non-opposites like *democratie - vredeproces* "democracy - peace process", as well as contrastive pairs that can be treated as opposites in certain contexts. For example, the pair *guerrilla strijder - paramilitair* "Guerilla fighter - paramilitary" is often antonymous in the newspaper contexts when the two are contrasted as "noble Robin Hoods" and criminals. Another example is the pair *boer - consument* "farmer - consumer", which is often contrastive in contexts related to the topic of agricultural policies, particularly price regulations. Such context-dependent pairs were unanimously discarded by the participants, as a result, based on the unanimous votes alone, the precision score for the 474 pairs found in the 5th iteration, was as low as 0.28.

The most difficult pairs for classification were the ones that did not receive unanimous votes either as opposites or as non-opposites. Among 38 pairs that did not receive unanimous votes as opposites were pairs *bewijs - vermoeden* "proof - presumption", *inhoud - stijl* "substance - style", *beperking - kracht* "limitation - strength", *Achilles hiel - kracht* "Achilles heel - power", *kantoor - woning* "office - residence", *religie - wetenschap* "religion - science". While none of these pairs have been discussed in literature on antonym classifications, such pairs are contrastive, not only in the newspaper contexts (like *Guerilla fighter - paramilitary*) but in a range of contexts. For example, *Achilles heel* is a metaphor for *weakness* (recall that originally in the mythodology the

heel of the Achille was his only vulnerable place), which is contrasted with *power*. The contrast between the *presumption* and *proof* is about the difference between the two: while *presumption* does not require any evidence, the *proof*, on the contrary, provides facts, which might refute the *presumption*. Thus, there is a contrast between absence and presence of proof. *Style* and *substance* are often contrasted with each other, for example, when technological innovations that either lack one or the other, or have both are discussed. All such examples seem to resemble indirect opposites, like *poor - wealthy* or *wet - parched*, in that their opposition is mediated by a third word and an extra step is required to recognize opposition.

This is similar with 63 pairs that did not receive unanimous votes as non-opposites. These pairs included *keuze - toeval* "choice - coincidence", *speler - supporter* "player - supporter", *ex-sporter - sporter* "former sportsman - sportsman", *fietser - voetganger* "cyclist - pedestrian", *inwoner - toerist* "resident - tourist", *architect - uitvoerder* "architect - subcontractor".

It is useful to recognize examples discussed above as good opposites, even if they are not canonical opposites. Although native speakers do not unanimously recognize such pairs as opposites, these opposites can be used for automatic identification of contrast. This also shows that it might be important to include context together with found pairs into classification tasks, because outside of the context opposites are often not recognized. If we treat such opposites and non-opposites by the majority vote as positive examples of opposites, the precision score based on the majority vote increases to 0.43, as opposed to the precision score of 0.28 based on the unanimous votes alone. This is still a low score in comparison to the results with dependency patterns for finding other relations.

Using manual classification, we can now evaluate the coverage of opposites in CORNETTO by examining how many of opposites by the majority vote are present in the resource and how many of them are marked as opposites. Out of the 141 pairs judged as opposites, 90.8% (128 pairs) contained both words in CORNETTO but only 7% of them (nine pairs) were linked as opposites. For example, pairs *kou - warmte* "coldness - warmth", *invoer - uitvoer* "import - export" were marked as opposites but pairs *emigrant - immigrant* "emigrant - immigrant", *haat - liefde* "hate - love" were not. Thus, our algorithm can improve the encoding of antonymy for at least 119 pairs that are already enlisted in CORNETTO.

With the manual classification of pairs, we can now examine how precision scores were influenced by the number of seeds. In particular, we address what happens with

Iteration	Found pairs	by majority vote	Opposites in CORNETTO	in CORNETTO & *MWB*	Precision scores based on judges	judges, CORNETTO & *MWB*
1	116	54.3% (63)	8.2% (9/110)	40.5% (47)	0.57	**0.6**
2	213	44.1% (94)	4.6% (9/194)	22.1% (47)	0.45	**0.47**
3	308	37% (114)	3.8% (10/264)	15.3% (47)	0.37	**0.39**
4	404	33.7% (136)	3.1% (11/350)	11.6% (47)	0.32	**0.35**
5	474	29.7% (141)	2.9% (11/372)	9.9% (47)	0.28	**0.29**

Table 6.11: Number of pairs with significant co-occurrence found per iteration, the percentage of how many were opposites according to the majority vote, the percentage of how many were present in CORNETTO and linked as opposites, and precision scores (based on unanimous votes; unanimous votes combined with the opposites in CORNETTO & *Mijnwoordenboek.nl*). The results are presented for the pairs found by means of 18 noun - noun seeds.

precision at each iteration as more pairs were used as seeds. The results are summarized in Table 6.11.

The results suggest that the best precision of 0.6 was achieved for the 116 pairs found during the first iteration when both lexical resources as well as unanimous votes were taken into account. Manual evaluation alone gave slightly lower results. Already at the second iteration the precision dropped from 0.6 to 0.47. This is 8% lower than the precision score achieved after the second iteration with adjective - adjective pairs (0.55). The precision at further iterations was steadily decreasing to 0.29 at the fifth iteration. In brief, the best precision was achieved when both manual evaluation as well as computational lexical resources were taken into account. *MWB* was particularly useful for noun - noun pairs as CORNETTO lacked many good pairs. Further, the highest precision score, comparable to those reported in other studies, was achieved only for the pairs found at the first iteration, suggesting that an increase in the number of seeds at further iterations introduced a lot of noise. In comparison with adjective - adjective seeds, dependency patterns identified with noun - noun seeds had lower precision.

Again, it is useful to assess the precision scores not only for the total number of found pairs but also for the top-k found pairs, as the most easily recognized opposites, according to manual classification, were among in the top results. Table 6.12 presents precision scores for the first 100, 150, 200 and 250 pairs found at each iteration, as well as the precision scores for the top-k pairs found by means of part-of-speech patterns, described in Chapter 5. First, this allows us to compare the precision of the algorithm across iterations, showing whether performance improves when the number of seeds

Top-k found pairs	Precision scores for pairs found with dependency patterns in iteration 1	iteration 2	iteration 3	iteration 4	iteration 5	PoS patterns -
100	0.62 (47)	**0.62** (48)	**0.62** (48)	**0.62** (48)	**0.62** (48)	0.61*
150	-	**0.57** (64)	**0.57** (64)	**0.57** (64)	**0.57** (64)	0.56*
200	-	0.47 (73)	0.47 (73)	0.47 (73)	0.47 (73)	**0.48** (72)
250	-	-	**0.45** (90)	**0.45** (90)	**0.45** (90)	0.43 (81)

Table 6.12: Precision scores for top-k pairs found in different iterations by means of dependency patterns (col. 2 - 6) and part-of-speech patterns (col. 7). The precision scores are based on the unanimous votes of the participants (shown in brackets), and opposites listed in CORNETTO and *MijnWoordenboek.nl*. All pairs were found by means of the same set of 18 noun - noun seeds. Scores with an asterisk are based on the average of the precision scores for 10 random samples because more than the top-k found pairs had a score of 1.

increases with newly found candidate opposites, as well as providing information as to how precision changes with an increasing number of pairs. Second, we can compare the precision of the algorithm based on the same number of top-k pairs found with dependency patterns as opposed to PoS patterns.

As is shown in Table 6.6, the precision of the algorithm with dependency patterns improved slightly in the second iteration, after which the results were consistent for all top-k pairs throughout consequent iterations. We reported similar results for the PoS patterns, namely, that there seems to be a finite number of good opposites after which the algorithm is not able to find new good pairs even with a larger number of seeds. Of course, by good opposites in this case we refer to pairs that are either well-established opposites or opposites that are easily recognized by the participants as such outside of the context. However, although opposites by the majority vote were not considered when estimating the precision scores, manual inspection showed that the number of opposites that did not receive unanimous votes was too small within top-k pairs across all iterations to significantly improve the precision. Thus, we can conclude that the increase of seeds in the consequent iterations does not help to find novel noun - noun opposites and the results do not improve after the second iteration.

6.4.2.2 Part-of-speech patterns versus dependency patterns for noun - noun pairs

In comparison with the performance of the PoS patterns, dependency patterns performed slightly better. For example, the precision scores for the top-100 pairs found

with dependency patterns in iterations 2 - 5 was 0.62 and the precision score for the top-100 pairs found with PoS patterns was 0.61. PoS patterns had a higher precision score than dependency patterns for the top-200 pairs. This result is similar to the findings of Tjong Kim Sang and Hofmann [2009], who also compared the performance of dependency patterns and PoS patterns for finding hyponyms and found that dependency patterns outperformed PoS patterns, when the corpus used was of the same size. They argued that adding more data would ensure higher precision of PoS patterns, concluding that PoS patterns perform as well as dependency patterns.

Another useful way of comparing the performance of dependency patterns with PoS patterns is by looking at the overlap between identified opposites found by means of both pattern types as well as only one of the pattern types. This is useful for two reasons. First, given that the two methods differed in the number of seeds and iterations different pattern types had, knowing the overlap can help to identify how similar the range of found patterns is. If the overlap between found patterns is large, then the differences in the precision of the algorithms can be improved by, for example, restricting the number of possible candidate pairs (as this was not done in the case of PoS patterns). A small overlap in found opposites, on the other hand, will indicate the possible limitations of the methods. For example, it might be that the total number of dependency patterns was too limited in our algorithm, restricting the possible range of candidate pairs. Because of that, the algorithm did not find any new opposites after the third iteration.

We compared two sets of found opposites. First, the overlap between the top-200 pairs found with PoS patterns were compared with the top-200 pairs found with dependency patterns in the last iteration. Since the opposites identified in each iteration were the same, this gave us a clear comparison with the top-200 pairs found with PoS patterns that were among the highest scored pairs. Second, the overlap between all opposites found with PoS patterns were compared with all opposites found in one of the five iterations with dependency patterns. The opposites consisted of pairs that were classified as opposites by at least two participants or were marked as opposites in one or both lexical resources. This illustrated the extent of the range of opposites found with different pattern types.

Among top-200 pairs found with PoS patterns, 95 pairs were judged as opposites by the majority vote or were opposites according to the lexical resources. Among the top-200 pairs found with dependency patterns 92 pairs were opposites according to the same classifications. Only 25 pairs were present in both sets. These opposites

included pairs *vijand - vriend* “enemy - friend”, *winter - zomer* “winter - summer”, *man - vrouw* “man - woman”, *overeenkomst - verschil* “resemblance - difference”, *roker - niet-roker* “smoker - non-smoker” and others. The other 74% of 95 opposites found with PoS patterns and 73% of opposites found with dependency patterns did not overlap with each other. Thus, dependency patterns and part-of-speech patterns find different opposites in the top-results.

There was an apparent striking difference between pairs only found with one of the two pattern types. In particular, among top opposites found with PoS patterns, many were opposites of kinship, for example, *oma - opa* “grandfather - grandmother”, *broer - zus* “brother - sister”, *moeder - vader* “mother - father”, and relational opposites (that is opposites that describe the same relationship with the two objects being reversed), for example, *docent - student* / *docent - leerling* / *leerling - leraar* “teacher - student”, *moeder* / *vader - zoon* “mother / father - son”, *kind - vader* “child - father” and so on. Opposites found with PoS patterns were also expressing contrast between more concrete nouns (for example, *acteur - actrice* “actor - actress”, *huurder - koper* “tenant - buyer”, *eigenaar - huurder* “owner - tenant”, *mens - dier* “human being - animal”) than opposites found with dependency patterns (for example, pairs *religie - wetenschap* “religion - science”, *tegenstander - voorstander* “opponent - supporter”, *bijval - kritiek* “support - criticism”, *beperking - kracht* “limitation - power”, *kracht - zwakte* “strength - weakness”, *Achilleshiel - kracht* “Achilles heel - power”. This points out that the most productive PoS patterns were qualitatively different from the most productive dependency patterns, finding different kinds of opposition.

In other words, the main difference in the top results found with the two pattern types is *not* the number of found opposites (both find approximately the same number of opposites according to the majority vote and lexical resources) but it is the kind of antonymous relationships they find.

If we look at the overlap between all opposites found with PoS patterns (407 pairs) and opposites found with dependency patterns (146 pairs), we can see that 67% of pairs found with dependency patterns are also found with PoS patterns (among pairs with the scoring $\geq$0.9). Thus, PoS patterns find the majority of opposites extracted by means of dependency patterns. Among 33% of pairs that were found only with dependency patterns were opposites *coalitie fractie - oppositie partij* “coalition party - opposition party”, *tegenstander - vriend* “opponent - friend”, *verarming - verrijking* “impoverishment - enrichment”, *vernieuwer - volger* “innovator - follower”, and others. Because there was no limit on the number of candidate opposites found with PoS

patterns, it is not useful to analyse pairs found only with PoS patterns and missed with dependency patterns as it can be that dependency patterns would also find them if the number of possible candidate opposites was not limited. However, the fact that PoS patterns identified many opposites found with dependency patterns suggests that the algorithm described in Chapter 5 is able to identify useful surface patterns that find the more abstract opposites found also by means of dependency patterns. However, such patterns are not among the most productive surface patterns.

In relation to previous studies on automatic relation extraction that used dependency patterns and PoS patterns for finding meronyms and hyponyms expressed by nouns, this is the first work that shows that at least for antonymy, a choice between pattern types (surface PoS patterns versus dependency patterns) needs to be based not only on such factors as algorithm efficiency and recall but also on such factors as the types of opposites in question.

6.4.2.3 Dependency patterns acquired with noun - noun seeds

The top dependency patterns acquired with noun - noun seeds were similar to the generic patterns acquired with adjective - adjective seeds and included equivalents of strictly PoS patterns [<*ANT/Noun*> *but* <*ANT/Noun*>], [<*ANT/Noun*> *and* <*ANT/Noun*>], [<*ANT/Noun*> *as well as* <*ANT/Noun*>], [*more* <*ANT/Noun*> *than* <*ANT/-Noun*>]. While such patterns have high recall, they are more general than similar PoS patterns and as a consequence they find a lot of noise.

Recall that for hyponym - hypernym extraction Pantel and Pennacchiotti [2006] dealt with the low precision of generic patterns like [<*X/Noun*> *and* <*Y/Noun*>] by filtering out incorrect instances of noun - noun pairs using the Web. However, in comparison to the TwNC, their corpus was very small (approximately 6 million words). Ittoo and Bouma [2010] used the same approach as presented above for meronyms. They used much larger corpora of approximately 110 million words for Dutch (Wikipedia texts) and approximately 470 million words for English (Wikipedia texts). They showed that large corpora are sufficient for identification of the best dependency patterns, which are short but not too general and contain the right instances of the meronym relation.

Although we used a much larger corpus for Dutch (450 million words from the newspaper texts) than Ittoo and Bouma [2010], many dependency patterns acquired for finding opposites were too general and contained a lot of non-opposites. Interestingly, the best dependency patterns acquired by Ittoo and Bouma [2010] for finding

Set1 & Set2	Only in Set1	Only in Set2	Overlap	Total unique
6 and 12 seeds	117	123	390	630
6 and 18 seeds	128	139	379	646
12 and 18 seeds	38	43	475	556

Table 6.13: Number of pairs found with verb - verb seed sets of different sizes and the overlap between them.

meronyms contained words like *to contain*, *to comprise* and so on. This seems to suggest that there are simply better dependency patterns for meronyms than for opposites.

Among other dependency patterns found with noun - noun seeds were patterns that contained dependency relations between the Subject and the Object. Such patterns received lower automatic scores than the generic patterns discussed above. One of the most productive surface patterns [*between <ANT/Noun> and <ANT/Noun>*] was not found with dependency patterns.

6.4.3 Results for verb-verb seed pairs

In comparison with the results obtained with seeds expressed by adjectives and nouns, results found with verb - verb seeds of different sizes had the largest overlaps between found pairs. As is shown in Table 6.13, 390 pairs (77%) found with six verb - verb seeds were also among pairs found with 12 seeds and 379 pairs (75%) found with six seeds were also among pairs found with 18 seeds. In comparison, 2.6% of pairs found with six adjective seeds and 37.4% of pairs found with six noun seeds were also found with the sets of 12 seeds of the same part-of-speech category. The largest overlap was found again among pairs found with 12 and 18 seeds. Namely, 475 pairs were found by both sets. The majority of dependency patterns found with the smallest seed set were the same as dependency patterns found with the larger sets. They included patterns similar to surface patterns [*to <ANT/Verb> and to <ANT/Verb>*], [*neither <ANT/Verb> nor <ANT/Verb>*], [*<ANT/Verb> as well as <ANT/Verb>*] and others. Since the results found with the set of 18 seeds contained the largest number of pairs found in the other two sets, we will discuss the results for the set of 18 seeds in detail.

Number of found pairs	Pairs with significant co-occurrence
512	86.7% (444)

Table 6.14: Number of pairs with found by means of dependency patterns with 18 verb - verb seeds and the percentage of pairs that co-occurred in the TwNC significantly more often than would be expected by chance.

6.4.3.1 Pairs found with 18 verb - verb seeds

Out of the total 512 unique pairs found after the fifth iteration 86.7% (444 pairs) co-occurred with each other within a sentence significantly more often than would be expected by chance (see Table 6.14). This was the lowest number of significantly co-occurring pairs, suggesting that verb - verb pairs are less likely to occur with each other significantly often. Among pairs with significant co-occurrence were *lezen - schrijven* "to read - to write", *teleurstellen - verrassen* "to disappoint - to surprise", *chatten - mailen* "to chat - to mail", *fietsen - wandelen* "to cycle - to walk", *verhogen - verlagen* "to raise - to lower", *annuleren - uitstellen* "to cancel - to postpone", *veranderen - verdwijnen* "to change - to remove", *blokkeren - hinderen* "to block - to hinder". Among pairs that did not co-occur significantly often were *gaan - wentelen* "to go - to turn", *lenen - werken* "to lend - to work", *aanmelden - aanpassen* "to subscribe - to adjust" and others.

Next, we examined how many significantly co-occurring pairs were opposites according to CORNETTO and *Mijnwoordenboek.nl*. The results are presented in Table 6.15. CORNETTO contained 91% (408 pairs) of the pairs, and 20 of them (4.9%) were linked as opposites. In other words, out of 444 found pairs, only 4.5% are confirmed opposites according to CORNETTO. Three pairs (15%) were linked as opposites *asymmetrically*. For example, *toenemen* "to increase" was among antonym candidates of *afnemen* "to decrease" but not the other way around. Among symmetric pairs were opposites *verkor- ten - verlengen* "to shorten - to prolong ", *benadelen - bevoordelen* "to aggrieve - to favour", *blijven - gaan* "to stay - to go", *scheiden - trouwen* "to divorce - to marry", *doorgaan - stoppen* "to continue - to stop", etc. Again, found asymmetry is a result of inconsistent coverage of opposites in CORNETTO, found for opposites expressed by adjectives, nouns and verbs.

Among pairs that were present in CORNETTO but not linked as opposites were *bellen - sturen* "to call - to send", *bevestigen - ontkrachten* "to endorse - to invalidate", *schilderen - tekenen* "to paint - to draw", *benoemen - ontslaan* "to nominate - to

Pairs with significant co-occurrence	In Cornetto	In *MWB*	In either one or both
444	4.9% (20/408)	4% (18)	6.3% (28)

Table 6.15: Percentage of found pairs listed as opposites in CORNETTO (col. 2), in *Mijnwoordenboek.nl* (col. 3) or in both resources (col. 4). The second number in column 2 represents the total number of found pairs, which have both words present in CORNETTO.

Opposites		Non-opposites		Total
by majority	unanimously	by majority	unanimously	
17.6% (78)	53.8% (42)	82.4% (366)	86.6% (317)	444

Table 6.16: Percentage of significantly co-occurring pairs found with dependency patterns with 18 verb - verb seeds classified as opposites or non-opposites by three participants. Unanimous counts are included in the majority vote.

dismiss" and others.

MWB contained 18 opposites (6.3% of all pairs), ten of which were also found in CORNETTO. Among them were pairs *verhogen - verlagen* "to raise - to lower", *breken - maken* "to break - to make", *afwijzen - toelaten* "to refuse - to allow", *accepteren - afwijzen* "to accept - to reject", and others. When both resources were taken into consideration, only 6.3% (28 pairs) of all found pairs were linked in either of them as opposites. The low number of opposites might be the result of the low coverage of opposites expressed by verbs in the resources. We will address this later in the section.

Next, we conducted a 'Yes/No' classification task, in which three participants were asked to evaluate each pair as an opposite or non-opposite. Participants achieved a Fleiss's kappa score of 0.58, which was the lowest score among pairs found with seeds expressed by adjectives, nouns and verbs. In particular, the Fleiss's kappa score for adjective - adjective pairs was 0.76 and the Fleiss's kappa score for noun - noun pairs was 0.67. A low level of agreement between participants suggests that it was more difficult for participants to evaluate pairs expressed by verbs, than pairs expressed by adjectives and nouns.

As can be seen in Table 6.16, approximately 18% of found pairs were classified as opposites by at least two participants. In comparison, 36.8% of pairs found with adjective seeds and 29.7% of pairs found with noun seeds were judged as opposites by the majority vote. Fifty-four percent of those pairs (42 in total) received unanimous

votes. Such pairs included opposites *halen - brengen* “to get - to bring”, *bouwen - slopen* “to build - to destruct”, *verhuren - verkopen* “to rent - to buy”. Another 46.2% of pairs (36 in total) were judged as opposites by the majority vote. Among such pairs were *investeren - sparen* “to invest - to save”, *aanhouden - vrijlaten* “to detain - to release” and others.

More than 80% of found pairs (366 in total) were judged as non-opposites, with approximately 87% (317 pairs) unanimously judged as non-opposites. A higher number of unanimously judged non-opposites as opposed to pairs judged as opposites suggests that it was easier for participants to decide on the pairs they thought were non-opposites. Among such pairs were *dineren - lunchen* “to dine - to lunch”, *bewaren - herstellen* “to preserve - to restore”, *aanraken - zien* “to touch - to see”, *behouden - krijgen* “to keep - to receive” and others.

Such pairs are contrastive in many contexts, for example *to touch - to see* are contrastive when different types of perception are compared; most of the time such pairs were not recognized as opposites because they belong to multiple-member categories, like *senses*, for which more than two members can be compared with each other (consider verbs of senses *to touch*, *to see*, *to smell*, *to hear* and so on). Since the difference between multiple incompatibles and co-hyponyms is not always transparent, such pairs were not consistently, that is unanimously, judged by the participants as opposites or as non-opposites.

Another difficulty with classification of such contrastive pairs is that they are not always mutually exclusive. So, *to touch* and *to see* are contrastive in some contexts but not others. It seems that found candidate verbs were more likely to belong to multiple-member categories than adjectives and nouns, making classification more difficult, which is reflected in a lower Fleiss’s kappa score. For example, *to dine - to lunch* are in the same category as the verb *to breakfast* and possibly verbs like *to snack*, *to eat out* and so on.

Interestingly, similar pairs were present among pairs that were judged as non-opposites by the majority vote (13.4% or 49 pairs). For example, pairs *blijven - teruggaan* “to stay - to return”, *ophalen - versturen* “to pick up - to send”, *horen - zien* “to hear - to see”, *weggeven - weggooien* “to give away - to throw away” and others.

In summary, manual evaluation of verb - verb pairs was the most difficult task for the participants, leading to lower agreement among participants. One of the main difficulties of the evaluation seems to be due to the fact that many found verb pairs were not binary, and they were part of a multiple-member category. Given that in

Iteration	Found pairs	by majority vote	Opposites in CORNETTO	in CORNETTO & *MWB*	Precision scores based on judges	judges, CORNETTO & *MWB*
1	111	25.2% (28)	9.6% (10/104)	18% (20)	0.24	**0.25**
2	201	21.9% (44)	7.3% (14/192)	11.4% (23)	0.18	**0.2**
3	296	20.3% (60)	4.9% (14/283)	7.8% (23)	0.15	**0.18**
4	378	19.9% (75)	5% (18/360)	7.1% (27)	0.13	**0.16**
5	444	17.6% (78)	4.9% (20/408)	6.3% (28)	0.12	**0.15**

Table 6.17: Number of pairs with significant co-occurrence found per iteration, the percentage of how many were opposites according to the majority vote, the percentage of how many were present in CORNETTO and linked as opposites, and precision scores (based on unanimous votes; unanimous votes combined with the opposites in CORNETTO & *Mijnwoordenboek.nl*). The results are presented for the pairs found by means of 18 verb - verb seeds.

the training session prior to classification participants were shown only canonical verb - verb opposites like *to buy - to sell* and *to begin - to end*, participants preferred to classify non-binary pairs as non-opposites, especially given that the context in which the words were antonymous was not presented to them.

Note that significant co-occurrence helped to reduce non-opposites from the results for verb - verb pairs. In particular, when significant co-occurrence was used as a cue, the number of non-opposites was reduced by almost 17%. This shows that significant co-occurrence can successfully reduce noise in the results. Also the agreement score between participants was slightly higher for the pairs with significant co-occurrences (0.58) than for all found pairs (0.56).

Overall, the high number of unanimously judged non-opposites was reflected in the low precision score. Namely, based on the manual evaluation alone, the precision score for the algorithm after the fifth iteration was 0.12. To examine how the performance of the algorithm changed with the increase in the number of used seeds, we calculated precision scores for each iteration. As discussed above, participants often could not recognize opposites outside of the context, especially for pairs from multi-member categories. In order to improve the reliability of the coverage of opposites in the results, the precision scores were calculated based on the unanimously judged opposites as well as opposites found in CORNETTO and *MWB*. The results are presented in Table 6.17.

The best precision scores were achieved when manual classification, CORNETTO and *MWB* opposites were taken into account. Still, this did not lead to good precision, and the results remained very low, with the highest precision score of 0.25 for the 111

Top-k found pairs	Precision scores for pairs found with dependency patterns in iteration 1	iteration 2	iteration 3	iteration 4	iteration 5	PoS patterns -
100	0.24 (20)	0.24 (20)	0.26 (21)	0.27 (22)	0.27 (22)	**0.56**
150	-	0.23 (28)	0.25 (29)	0.24 (28)	0.24 (28)	**0.45**
200	-	0.2 (33)	0.2 (32)	0.2 (32)	0.2 (32)	**0.38**
250	-	-	0.18 (37)	0.19 (38)	0.19 (38)	-

Table 6.18: Precision scores for top-k pairs found in different iterations by means of dependency patterns (col. 2 - 6) and found by means of part-of-speech patterns (col. 7). The number of the unanimously judged opposites in each set is given in brackets. All pairs were found by means of the same set of 18 verb - verb seeds.

significantly co-occurring pairs found at the first iteration. This precision was lower than the precision scores for the pairs found after the fifth iteration with 18 adjective - adjective and noun - noun seed sets. In particular, the precision scores for the adjective pairs were 0.69 after iteration one and 0.38 after iteration five (based on 459 pairs) and the precision scores for the noun pairs were 0.6 after iteration one and 0.29 after iteration five (based on 474 pairs). Thus, the algorithm based on dependency patterns performed least well for finding opposites expressed by verbs.

As is shown in Table 6.18, the precision score was very low even when only top-100 pairs were considered. Namely, the precision score for the top-100 pairs found in iteration one was 0.24 and in iteration five it was 0.27. In comparison, the precision score for the top-100 pairs found with surface PoS patterns and the same set of 18 verb - verb seeds was 0.56, for the top-150 pairs, the precision score was 0.45 and for the top-200 pairs, it was 0.38. Thus, PoS patterns outperformed dependency patterns at finding opposites expressed by verbs.

6.4.3.2 Dependency patterns acquired with verb - verb seeds

Analysis of dependency patterns identified by means of verb - verb seeds is especially useful for understanding why the algorithm performed particularly weakly for this part-of-speech category. Surprisingly, the majority of dependency patterns identified by means of verb - verbs seeds were generic patterns like [*to* <*ANT/Verb*> *or* <*ANT/Verb*>], [<*ANT/Verb*> *but* <*ANT/Verb*>], [*more* <*ANT/Verb*> *than* <*ANT/Verb*>], [<*ANT/Verb*> *as well as* <*ANT/Verb*>], [*neither* <*ANT/Verb*> *nor* <*ANT/-Verb*>] and their derivatives. Such patterns are too general to find mostly opposites. This is an interesting finding in relation to the previous work on dependency patterns

in relation extraction. Particularly, it shows that syntactic information is useful for finding lexical semantic relations expressed by adjectives and nouns whereas in the case of verbs it can even hurt the results.

6.5 Discussion

In this chapter we presented an automatic method for finding opposites by means of dependency patterns. Recall that dependency patterns contain information about syntactic relations between words and, as a result, they abstract away from the linear ordering of words. Our goal was two-fold. First, we examined whether dependency patterns can successfully find opposites expressed by adjectives, nouns and verbs. Second, we wanted to know whether opposites found by dependency patterns are different from opposites found by means of surface PoS patterns.

In relation to the performance of dependency patterns, we found that they produce different results, depending on the syntactic category of seeds and candidate opposites. The best precision scores were achieved for the adjective - adjective seed sets and the lowest precision scores were found in the results with verb - verb seeds. And while the increase in the number of seeds in consequent iterations did not improve the recall, the precision scores for the top-k pairs improved at later iterations for adjective - adjective pairs but not for noun - noun and verb - verb pairs. This means that a small set of seeds is sufficient for identification of well-known opposites, which seem to be restricted in number by the corpus size and maybe also the genre. Once such pairs are extracted, the majority of novel candidate opposites found in the same pattern types are context-dependent pairs that are not always easily classified as opposites or non-opposites.

This is an interesting result because it shows that a much wider range of contrastive pairs than has been previously recognized, are found in productive patterns that contain well-established opposites. And any theoretical account on antonym classification needs to take these pairs into consideration, at least from the perspective of discourse functions of opposites related to the pattern types in which they are found.

However, speaking of discourse functions, recall that Jones [2002] studied co-occurrence of canonical opposites in textual patterns, which differ from dependency patterns. Interestingly, in comparison to the most productive surface PoS pattern types, top dependency patterns were different. Namely, while the majority of surface patterns used in previous experiments were very specific and relatively long (six elements long

on average), the best dependency patterns according to the algorithm were rather general, for example, equivalent to surface patterns [*X and Y*], [*X but Y*], [*X or Y*] and, therefore, noisy. Interestingly, only six seeds found specific dependency patterns but most of the pairs they extracted were not opposites. Adding more seeds resulted in finding very general patterns that extracted more opposites than specific dependency patterns. Nevertheless, generic dependency patterns also extracted a lot of noise.

Recall that an *Espresso*-like algorithm presented in this study is different from the approach taken with surface patterns in Chapters 4 and 5. In the original work of Pantel and Pennacchiotti [2006] generic patterns are at the core of their algorithm for finding hyponym-hypernym pairs. They dealt with the low precision of generic patterns by means of the Web. They showed that the generic patterns were an added value for the *Espresso*-algorithm but it was necessary to use the Web in order to filter out noisy instances from the results. Note that their corpus was relatively small, and consisted of approximately six million words only. Ittoo and Bouma [2010] adapted the same algorithm for finding meronyms but since they had a much larger corpus (approximately 450 million words) they argued that using the Web was not necessary. They successfully found meronyms arguing that given enough data the Web is not needed.

We used the same algorithm as proposed in Ittoo and Bouma [2010] on the same newspaper corpus for Dutch (TwNC). However, our results show that when a lexical semantic relation is not defined by any specific (frequently occurring) patterns, a large corpus is not enough for filtering out noise from the results. Most of the dependency patterns acquired by Ittoo and Bouma for finding meronyms were of the type [<*X/Noun*> *contains / includes / comprises* <*Y/Noun*>] which are very likely to indicate meronyms. We, on the other hand, found rather generic patterns like [<*ANT/Noun*> *but* <*ANT/Noun*>]. Given that the strictly PoS patterns gave better results (especially for nouns and verbs), our findings are in line with the work of Tjong Kim Sang and Hofmann [2009] who suggest that strictly PoS patterns are as good as dependency patterns (they were interested in hypernym-hyponym pairs expressed by nouns). We show that strictly PoS patterns are as good as dependency patterns for finding opposites expressed by adjectives and nouns and better than dependency patterns for finding opposites expressed by verbs. Since shallow parsing is a fast and efficient preprocessing step that can be applied to a vast amount of data, we conclude that strictly PoS patterns can be used more productively for finding opposites than dependency patterns.

Going back to the differences between types of PoS and dependency patterns, it might seem that because the types were so different, the kinds of opposites they found

were also qualitatively different from each other. However, this is not the case. On the one hand, PoS patterns are more specific and reflect very productive linear patterns like [*difference between* <*ANT*> *and* <*ANT*>], similar to the specific pattern types like [*X contains Y*] that are usually used for finding meronyms. On the other hand, dependency patterns are more general and find opposition between abstract nouns more frequently than PoS patterns, making it more likely to identify abstract concepts, not frequently compared in productive surface patterns. With both pattern types significant co-occurrence within a sentence is a strong cue for antonymy, although the distance between opposites can vary depending on such factors as how frequently the pair of opposites is found in the genre of the corpus, how conventionalized it is as opposites and how context-dependent it is. However, opposites identified with dependency patterns were also found with surface PoS patterns among pairs with lower automatic scoring. This indicates that although those pairs are less frequently found together they do co-occur in close proximity to each other in productive surface pattern types.

We do find large differences in the results for opposites expressed by adjectives, nouns and verbs. For example, dependency patterns performed extremely poorly with verb - verb pairs. First of all, because the lowest number of significantly co-occurring pairs was found in the results with verb - verb seeds, it can be that, unlike adjectives and nouns, opposites expressed by verbs are less likely to co-occur with each other within a sentence than opposites expressed by nouns and adjectives. In fact, Hielkema (2007) argues that opposites expressed by verbs tend to co-occur with each other not within a sentence but rather within several paragraphs of the same text. This explains why pattern-based methods perform least well with verb - verb seeds. However, PoS patterns performed better at finding opposites, leading to higher precision for the top-k pairs. Thus, the difference in the performance of dependency patterns as opposed to surface PoS patterns might lie in the kind of textual functions of verbal opposites.

The usefulness of finding opposites automatically can be illustrated by looking at examples of the inconsistencies of manual evaluation of the results. While dependency patterns found many established opposites, they also extracted many non-typical, often context-dependent opposites which participants often failed to recognize as contrastive or incompatible. For example, adjective - adjective pairs *anderhalf - half* “one and a half - half”, *neutraal - positief* “neutral - positive”, *leerzaam - vermakelijk* “informative - entertaining” were not recognized as opposites but (1) they exhibit the same behaviour in the corpus as well-established canonical pairs, leading to their high automatic ranking; (2) they co-occur with each other within a sentence significantly more

often than is expected by chance - a prerequisite used for separating non-antonymous pairs from the results; (3) they indicate contrast or incompatibility in specific contexts.

Such examples have been neglected in theoretical classifications of opposites. Finding opposites automatically is therefore a reliable, methodologically-sound way of finding new classes of those opposites that are not recognized as such by the native speakers when the context is not provided. The fact that in some cases all three participants dismissed context-dependent opposites shows the limitations of approaches to antonymy that are based on the intuition of native speakers. As has been discussed in the previous section, manual evaluation of automatically found pairs seems to reflect a spectrum of the degrees of antonymy. Namely, easily recognized, canonical opposites received unanimous votes (for example, the pair *hot - cold*). Opposites that do not receive unanimous votes contain multiple incompatibles, that is non-binary opposites (for example, the pair *to hear - to see*). The next class of opposites comprises pairs that are strongly context-dependent. For example, pairs judged as non-opposites by majority vote include *white - red* (contrastive in relation to wine), *boer - consument* "farmer - customer" (contrastive in the context of production and consumption). Our algorithm finds all types of these opposites because their corpus profiles are very similar. In other words, all these pairs co-occur significantly often in the same contexts (or patterns). Previous studies of antonymy used significant co-occurrence, types of patterns and native speakers' intuition to classify well-recognized pairs of opposites. Our results highlight that these properties are also characteristic of less typical and often counter-intuitive opposites that have not yet been studied.

CHAPTER 7

Discussion

In this dissertation, we addressed two central topics: how pattern-based methods can be applied to antonym harvesting and how automatically found opposites correspond to the theoretical classifications on antonymy proposed in theoretical linguistics.

In relation to the first topic, we studied three different pattern-based methods for automatic extraction of opposites. The first kind of patterns was based on the surface structure of a sentence and did not require any annotation of the corpus. The second kind of patterns was based on the surface structure of a sentence but it required part-of-speech information about the candidate pairs, which had to belong to the same syntactic category as the seeds. The third kind of patterns did not rely on the surface structure of the sentence but required full parsing as these patterns contained information about syntactic relations between words. We showed that textual part-of-speech patterns (the second kind) and dependency patterns (the third kind) outperformed surface patterns (the first kind). There are, however, important differences between the two best pattern kinds that will be discussed in detail in Section 7.1.

In relation to the second topic, we explored the types of opposites found in automatically identified surface patterns as compared to the types of opposites described in the existing theoretical classifications. We showed that the range of automatically

found opposites and contrastive pairs goes beyond the limited number of the examples of canonical and non-canonical opposites commonly discussed in the literature. We showed that automatic methods are capable of identifying not only canonical but also context-dependent opposites that are unlikely to be recognized as opposites when encountered without any context. Interestingly, we also found that there are differences as to the types of the most productive patterns, depending on the kind of the pattern-based method we used. Moreover, we found that the same opposites co-occurred in different kinds of patterns, suggesting that a method that studies the discourse functions of opposites using surface patterns alone, in particular Jones [2002], might be limited. The details of these findings will be discussed in Section 7.2. Because this is the first study that explores the similarities and differences between manually and automatically identified opposites, the large contribution of our work is a better understanding of the types of opposites found in corpora.

7.1 The best performing method

Part-of-speech patterns versus dependency patterns. Surface part-of-speech patterns found the largest overall number of classified opposites based on manual classification and computational lexical resources CORNETTO and *Mijnwoordenboek.nl*. Part-of-speech patterns found the largest number of opposites for each of the three syntactic categories. This kind of patterns also had the highest recall for all three syntactic categories, returning the largest number of candidate pairs with each seed set. This was an unexpected result, as we thought that adding part-of-speech information would restrict the possible number of the returned results. However, the results suggest that sentential co-occurrence of opposites in surface patterns is strong not only among adjective - adjective but also among noun - noun and verb - verb pairs, in accordance with the proposal of Fellbaum [1995].

Recall that the method that used dependency patterns returned 100 found pairs, adding an extra 100 found pairs at each consecutive iteration to the already found opposites treated as new seeds. As a result, although dependency patterns found fewer opposites overall, the opposites identified by the dependency patterns were among the top-k results, leading to higher precision scores. Given that opposites found by dependency patterns were also found by part-of-speech patterns suggest that part-of-speech information is sufficient for finding a wide range of opposites. However, part-of-speech patterns found a lot of noise, and before part-of-speech patterns can be used for reliable

antonym harvesting, a more effective way of filtering out noise from the results have to be found.

One of the potential factors that can improve the precision of the part-of-speech patterns is the size of the corpus. As Tjong Kim Sang and Hofmann [2009] argue, when they used a 20% larger corpus, part-of-speech patterns performed as well as dependency patterns at finding hypernym-hyponyms in Dutch. As presented in chapter 4, section 4.6.3, our results showed that with strictly textual patterns more data gave similar precision and recall with fewer seeds than less data with more seeds. What this suggests is that a pattern-based method can be used on a relatively small corpus, for example, in comparison to the World Wide Web. At the same time, using a larger data set when available might yield higher precision. Given that part-of-speech parsing is much faster to perform than the full parsing required for generation of dependency patterns, this method can and should be tested in the future.

One interesting difference between our results and the results described in Tjong Kim Sang and Hofmann [2009] with hyponyms. Tjong Kim Sang and Hofmann report that their pattern types within dependency patterns and part-of-speech patterns were similar, the pattern types we found among the most productive part-of-speech patterns and dependency patterns were different. For example, there was no comparative pattern type [*the difference between* <*ANT*> *and* <*ANT*>] found with dependency patterns. Overall, dependency patterns were more generic than part-of-speech patterns. This means that opposites expressed by all three part-of-speech categories co-occur in a wide range of pattern types with varying surface distance between the two words in a pair. Some of the pattern types can be identified and manually recognized as contrastive, whereas others are too general and cannot be manually recognized as productive patterns based on researcher's intuition alone. In contrast, other lexical and semantic relationships, for example, hyponymy, can be automatically identified by similar pattern types. It is important to take this into consideration when studying discourse functions of opposites Jones [2002], Jones et al. [2007], Willners and Paradis [2010], as this highlights that the functions of opposites in discourse are neither manifested, nor limited by surface pattern types in which canonical opposites can be found.

Also, using larger seed sets can positively affect the precision. For example, dependency patterns found more opposites at each iteration when found opposites were added as new seeds.

As has already been mentioned, significant co-occurrence is not a sufficient cue for eliminating noise from the results. A more diverse approach that uses more than one

method for antonym validation might help to separate opposites from non-opposites. Recall that distributional methods, discussed in detail in chapter 2, section 2.4.2, can successfully identify opposites in large corpora. However, the limitation of distributional methods is that they cannot separate opposites from synonyms, which also tend to share similar contexts. Lobanova et al. [2010] proposed to use automatically found opposites to eliminate noise, that is erroneously found opposites, from the results of a distributional method aimed at finding synonyms. Their approach did not significantly improve the precision, mostly because the number of found opposites was small. Nevertheless, it seems promising to use the results from a distributional method for validation of good opposites found by means of a pattern-based method.

7.2 *Automatically found opposites*

All three pattern-based methods found many opposites, but part-of-speech patterns and dependency patterns outperformed strictly textual patterns in both the number of found opposites expressed by all three part-of-speech categories and the types of opposites they found. In particular, part-of-speech patterns and dependency patterns found a wide range of non-canonical opposites, showing that automatic methods for antonym harvesting provide a useful and powerful means of identification of a wide range of opposites. Both pattern types found pairs like *gedwongen - vrijwillig* “compulsory - voluntary”, *verarming - verrijking* “impoverishment - enrichment” and context-dependent pairs like *duif - havik* “dove - hawk” (contrastive in the non-literal meaning to compare peaceful and aggressive people), *groen - zwart* “green - black” (contrastive in the context of coffee and tea blends) and *internationaal - Nederlands* “international - Dutch” (contrastive in social and political texts when local and international policies are compared). The latter pairs were often not recognized by judges as opposites in the evaluation tasks. This suggests that automatically identified patterns are able to find very atypical opposites that researchers are unlikely to come up with in corpus-based research like the study of Jones [2002]. Automatically found patterns also provide means to study the differences between canonical and non-canonical opposites and their functions in discourse by looking at the types and the number of patterns in which they co-occur in newspaper texts.

7.2.1 Antonym canonicity

As has been discussed in Chapter 2, previous studies, in particular Jones et al. [2007], suggest that it is possible to use patterns to determine antonym canonicity (see Section 2.3.4 for details). In particular, Jones and colleagues argue that the range of patterns in which a pair occurs, or its "breadth of co-occurrence", is a strong indicator of its canonicity. Based on their results, the authors report that canonical opposites tend to co-occur in ten or more patterns whereas non-canonical opposites co-occur in fewer patterns. Note that by ten patterns the authors meant five distinct pattern types with two possible orderings of the opposites in each pattern. For example, the pattern variations [*between* <*ANT*> *and* <*ANT*>] and [*between* <*ANT*> *and* <*ANT*>] were treated as two patterns. It was not possible for us to apply the same approach because Jones et al. [2007] used one seed word to see how often it would retrieve the other seed word and vice versa in manually preselected patterns. But this was not necessary in our case because we automated the step of pattern identification and validation and used both words from seed pairs together, disregarding their ordering. In this way we were able to find many more contrastive patterns that retrieved a wide range of opposites that were not previously studied, including the work of Jones et al. [2007].

The advantage of our approach over previous work is that we did not limit the range of possible candidate pairs by constraining the types of patterns in which they can be found. At the same time, our results allow us to examine whether the same tendency of co-occurring in more patterns is found with opposites found in automatically identified patterns. This can be studied by looking at the automatic scoring of found pairs and the number and the types of patterns in which they were found.

Our results show that canonical and non-canonical opposites were equally likely to be found in productive patterns, receiving equally high top scores. The same finding holds for all pattern types: strictly textual patterns, surface part-of-speech patterns and dependency patterns.

To illustrate this, consider the following example. The canonical adjective - adjective opposites *nieuw* - *oud* "new - old" were found in 12,486 surface part-of-speech patterns and received the highest automatic scoring of one.[1] Both words were also found with other candidates, forming other good pairs of opposites with the highest automatic scores. For example, the word *new* was also found with the opposites:

[1]Recall that each candidate pair received an automatic score between 0 and 1 with 0 suggesting that a pair is not likely to be antonymous and with 1 suggesting that a pair is very likely to be antonymous. This score reflected the number and the goodness of the patterns in which a pair was found.

bestaand “existing” (in 173 patterns);
huidig “current” (in 45 patterns);
traditioneel “traditional” (in 50 patterns);
gevestigd “established” (in 26 patterns);
gebruikt “used” (in 24 patterns);
bekend “familiar” (in 17 patterns);
klassiek “classical” (in 26 patterns);
vorig “previous” (in 19 patterns);
vertrouwd “familiar” (in 21 patterns);
tweedehands “second-hand” (in 15 patterns).

And the word *old* was also found in combination with the following opposites:
modern “modern” (in 154 patterns);
klein “small” in the sense of “young” (in 27 patterns);
recent “recent”(in 50 patterns);
jong “young” (in 7823 patterns);
hedendaags “contemporary” (in 15 patterns);
vers “fresh” (in 24 patterns).

Although the canonical opposites were found more frequently than any of the above combinations, it seems that the difference in the number of patterns reflects the high frequency lexical pairing of the words *new* and *old* rather than their canonicity. Less frequent pairings still co-occurred in patterns quite often, ranging between 173 and 15 patterns. It is difficult to say whether the number of patterns in this range played any role, since some pairs were found in fewer patterns with high automatic scoring and others co-occurred in more patterns but with lower automatic scoring.

Interestingly, the types of patterns as opposed to their number did not play the same role as has been suggested in Jones et al. [2007]. Recall, that Jones and colleagues argue that co-occurrence in more pattern types and reciprocity of opposites are strong indicators of their canonicity. We found that the pattern types like [<*ANT*> *and/or* <*ANT*>], [<*ANT*> *as well as* <*ANT*>], [*between* <*ANT*> *and* <*ANT*>] were popular with all found opposites listed above. On the other hand, the pattern of incompatibility [<*ANT*> *versus* <*ANT*>] was very infrequent even with the canonical opposites *new - old*, in which they were found three times. This shows that manually selected patterns are not sufficient for studying antonym canonicity because some of them are infrequent in natural language.

In relation to manual classification, most of the pairs listed above were judged by

the participants as unanimous opposites, even though the participants did not always recognize non-typical opposites. A few pairs, namely, *nieuw - bestaand* "new - existing" found in 173 patterns and *nieuw - huidig* "new - current" found in 45 patterns, were judged by the participants as non-opposites by the majority vote. Three other pairs, namely, *nieuw - gebruikt* "new - used" found in 24 patterns, *nieuw - vorig* "new - previous" found in 19 patterns and *oud - vers* "old - fresh" found in 24 patterns, were classified as opposites by the majority vote. And the pair *oud - klein* "old - small" found in 27 patterns was unanimously discarded as non-opposites. It seems that rather than canonicity, the difference in the classification of these pairs is related to their context-dependency. Regardless of the patterns in which they were found, the pairs *old - small* and *new - current* seem to require context to be recognized by the participants as opposites whereas the pairs *new - familiar*, *new - second-hand* or *new - classic* do not. In other words, because the classification task did not provide any context, participants did not recognize opposites among pairs with less-frequent senses.

In relation to the psycholinguistic studies on antonymy that use elicitation tasks, our results imply that most of the automatically found opposites are likely to be dismissed because the participants are likely to come up with the most frequent highly associated opposites only. Unlike theoretical and psycholinguistic studies of antonymy, corpus-driven studies of opposites are able to deal with words that have multiple opposites. For example, among the opposites listed above, the part-of-speech patterns also found other pairs of opposites, 35 in total, such as *traditional - modern*, *classic - contemporary*, *fresh - dried*, *old-fashioned - contemporary* and so on. This highlights the advantages of using automatic antonym harvesting techniques for finding a wide range of opposites and the importance of using data-driven approaches to studying antonymy. Another reason to question the relevance of canonicity in the study of antonymy is the fact that canonicity is always discussed in relation to opposites expressed by adjectives. In particular, psycholinguistic experiments focus solely on the adjectives. Our results show that opposites expressed by nouns are as common if not more common than opposites expressed by adjectives. These findings are in line with the earlier results of Lobanova et al. [2010] who show that noun - noun opposites are found in the Dutch newspaper texts more frequently than opposites expressed by adjectives. In contrast, we found that opposites expressed by verbs are infrequent and they are the hardest to evaluate.

Interestingly, with nominal and verbal opposites, canonical opposites we used as seeds did not exhibit different pattern behaviour from found pairs. In fact, sometimes

the seed pairs were found in fewer patterns than novel candidate pairs. For example, the seed pair *top - bodem* "top - bottom" was found 56 times, the seed pair *slagen - mislukken* "to succeed - to fail" was found 83 times. In contrast, the novel noun-noun pair *kerk - staat* "church - state" was found 668 times and the novel verb - verb pair *wonen - werken* "to live - to work" was found 166 times.

Given that we do not find behavioural differences in the corpus between canonical and non-canonical opposites, instead of dividing opposites into canonical and non-canonical, we propose that opposites differ as to the degrees of antonymy originally suggested by Mohammad and Turney [2010]), and discussed in details in Section 2.4.2 of Chapter 2. This distinction is also reflected in the differences in the evaluation of found pairs by the participants.

7.2.2 Degrees of antonymy

As has been discussed in section 2.4.2, chapter 2, the degrees of antonymy are said to reflect the differences between opposites that are perceived by the participants of psycholinguistic experiments as 'better' than other opposites. Although the idea of antonym canonicity is closely related to degrees of antonymy in that opposites with high degree of antonymy can also be canonical, there is a very important difference between the two. Namely, even though there is no clear-cut way of separating canonical opposites from non-canonical opposites, the underlying assumption is that a pair is always either canonical or not. The degrees of antonymy, on the other hand, allow for variation of 'goodness' of opposites, which is relative and depends on the set of given opposites. For example, when *thin* is contrasted with *thick* and *chubby*, the word *thick* might be selected as a 'better' opposite than *chubby*. But in the candidate set *chubby* and *plump*, the word *chubby* might be selected as a 'better' opposite of *thin* than the word *plump*. These differences in the degrees of antonymy are based on the co-occurrences of candidate pairs in corpora.

Although we asked participants to evaluate found pairs as opposites or non-opposites, we found that the differences between found pairs were more fine-grained than the dichotomy associated with canonicity. In particular, it seems that found opposites had different degrees of antonymy, reflected in whether they were judged as opposites by all three participants or only by the majority. Even pairs that were judged as non-opposites by the majority vote were more contrastive than pairs judged as non-opposites by all three participants.

For example, the pairs *man - vrouw* “man - woman” and “husband - wife” were judged as opposites by all three participants, the pair *echtgenoot - vrouw* “spouse - wife” was judged as opposites by the majority vote. This shows that opposites that share many contexts across different topics and genres are more easily recognized by the participants as opposites, whereas pairs that share few contrastive contexts and only in certain domains are often not recognized by the participants as opposites or they are felt to be less ‘good’. Our findings support the results of the elicitation experiment of Paradis et al. [2009], who show that for some stimulus opposites it was not possible to identify one best opposite and participants named more than one equally good candidate.

The pairs *huidig - toekomstig* “current - future”, *vandaag - morgen* “today - tomorrow” and *goud - zilver* “gold - silver” were among pairs judged as non-opposites by the majority vote. And the pairs *klein - middelgroot* “small - middle”, *migrant - Nederlander* “migrant - Dutch” and *Arabisch - Westers* “Arabic - Western” were judged as unanimous non-opposites, although in the newspaper texts the latter are often used to contrast different cultures. The fact that none of these pairs were judged as opposites at least by the majority vote shows that they have lower degrees of antonymy than pairs like *man - vrouw* “man - woman”. Moreover, the pairs above that were unanimously dismissed as non-opposites, can also be contrastive in certain domains like political texts. In the future, it needs to be established whether, and how, adding the context for candidate opposites in the evaluation tasks will affect the way pairs are classified as opposites or non-opposites with more pairs being recognized as opposites.

Surprisingly, so far, the degrees of antonymy have only been discussed in computational work on antonym harvesting. Instead, theoretical linguists tend to rely on the dichotomy between canonical and non-canonical pairs. However, the degrees of antonymy nicely explain the differences in the perception of the goodness of opposites in manual evaluation experiments, whereas canonicity does not. Therefore, such an approach offers a promising direction in the further study of antonymy. For example, the concept of the degrees of antonymy can be studied by conducting psycholinguistic experiments with automatically found opposites, examining how the ’goodness’ of opposites changes depending on the set of candidates and the context, in which the opposites are shown.

The degrees of antonymy seem to be also more useful when using opposites in Natural Language Processing applications that rely on automatically found opposites. For example, it is more useful to identify non-typical opposites that express contrast in a

certain specific context or a specific register than to rely on a set of well-established thoroughly studied and classified canonical opposites. Thus, while the concept of canonicity seems to be an artefact of theoretical studies, the concept of the degrees of antonymy is a more empirically-grounded finding.

7.2.3 Theoretical classifications and automatically found opposites

Seed sets expressed by all three part-of-speech categories identified good opposites, although verb - verb seeds found the least number of opposites. The majority of found pairs that were classified as opposites were expressed by adjectives and nouns. While finding many opposites expressed by adjectives was not surprising, as adjective - adjective opposites are the best studied antonymous pairs, similar to the findings of Lobanova et al. [2010], the seed sets expressed by nouns found unexpectedly many nouns, suggesting that opposites expressed by nouns are as likely to co-occur in contrastive patterns as adjectives. As these pairs fall under the category of non-gradable opposites, it is particularly interesting to know that there are many more non-gradable opposites than has been previously recognized, and that these opposites tend to co-occur in the same productive patterns as well-established opposites expressed by adjectives. Moreover, given that in strictly textual patterns noun - noun opposites were also found with adjective - adjective seeds, and vice versa, it seems that opposites expressed by adjectives and nouns share common contrastive contexts in which they can be found automatically.

It is useful to look in more detail at the types of automatically found opposites expressed by different part-of-speech categories and how they fall under the established types of opposites. Recall that the two main distinctions between opposites and non-opposites that have been proposed by different theories on antonymy are based on whether a candidate pair is gradable and whether it is binary (see section 2.2 chapter 2 for details).

7.2.3.1 Opposites, gradability and binary dichotomy

Among the top results found with adjective - adjective seeds and part-of-speech patterns we found canonical gradable opposites like *sterk - zwak* “strong - weak”, *dun - dik* “thin - thick” and *gezond - ziek* “healthy - sick”, *conservatief - progressief* “conservative - progressive”. However, these pairs contributed a small part of adjective - adjective opposites found in the newspaper corpus. Most of the adjective - adjective

opposites were non-gradable pairs, including mutually-exclusive complementaries like *intern - extern* “internal - external”, *analoog - digitaal* “analogue - digital”, *lichamelijk - geestelijk* “bodily - mental”. Both types of pairs were unanimously recognized as opposites by judges. The fact that gradable and non-gradable opposites were found in similar patterns with similar frequency, both receiving the highest automatic scoring and that they were classified as opposites by all three judges suggests that gradable and non-gradable adjective - adjective pairs are equally good opposites. This means that the ‘goodness’ of antonymy should not be based on whether a pair is gradable or not.

Further, the participants also unanimously classified the pair *westers - oosters* “Western - Eastern” as opposites, although strictly speaking this is a non-binary pair of mutual incompatibles that refer to cardinal points: *north - east - south - west*. Because they are not binary, most of the theoretical linguists, including Cruse [1986], do not recognize them as opposites. But, similar to the binary opposites discussed above, this pair shares many contrastive contexts, for example, referring to the extremes of the Eastern and Western cultures among other examples, and the participants recognize it as antonymous even when no additional contrastive contexts are provided.

Recall that with the sets of multiple incompatibles with four members, for example, the seasons of the year or the set *man - woman - girl - boy*, it is said that each member is in opposition with two other members. While our algorithm found different combinations of the pairs all with high automatic scoring, for example, *northern - southern*, *western - southern*, and *eastern - southern*, only the directional extremes, such as *northern - southern* and *eastern - western* were unanimously judged as opposites. What this shows is that participants were more inclined to recognize more contextually relevant opposites. This, in turn, affects the evaluation of the performance of the algorithm that found naturally co-occurring contrastive pairs.

The role of the context in antonymy and how it affects human intuition about opposites can be illustrated on the following examples. Because we used a newspaper corpus, many articles contained stories about local and international events, often comparing the two. The patterns found the pairs *binnenlands - buitenlands* “domestic - foreign” and *buitenlands - nationaal* “foreign - national”; both of them were unanimously judged as opposites. The patterns also found the indirect antonym pair *buitenlands - nederlands* “foreign - Dutch”, where *Dutch* stands for the opposite *local*. The participants did not have any difficulties recognizing these as opposites by the majority vote. However, they failed to recognize the same opposition for similar pairs that received equally high automatic scoring: *buitenlands* “foreign” - *frans* “French” / *rus-*

sisch "Russian" / *iraaks* "Iraqi" / *chinees* "Chinese" / *palestijns* "Palestinian" / *italiaans* "Italian" / *duits* "German" and others. All these pairs were unanimously discarded as non-opposites. This again shows that automatic methods for finding opposites are unbiased towards researcher's intuition, being able to identify context-dependent opposites that are otherwise dismissed.

Note, that the examples given above do not fall under any of the proposed categorizations of opposites. They are not gradable, not binary, but what they all have in common is their high co-occurrence in contrastive patterns. These are examples of contrastive pairs that are more or less antonymous depending on the context. Following Murphy [2003], who suggests that opposites should be viewed as a context-dependent phenomenon, we support this claim, viewing contrastive sets above as context-dependent opposites.

7.2.3.2 Manual classification of found pairs

Because our methods found a large number of non-canonical context-dependent opposites, the task of the classification of the results proved to be difficult. May be because the participants had to evaluate many pairs at once and no context was provided, it was difficult for the participants to classify non-conventional opposites. Some participants also failed to recognize even well-established opposites. Nevertheless, the agreement among participants was substantially high, especially for pairs expressed by adjectives.

7.3 Summary

This dissertation presented research on opposites, in particular, we explored three pattern-based methods for automatic extraction of opposites from large newspaper text collections. We compared the results from algorithms that differed as to the amount of syntactic information they required. In the first study, we examined the performance of automatically generated strictly textual patterns, that is patterns that do not contain any syntactic information about the target pairs, for example, [*difference between <ANT> and <ANT> countries*]. In the second study, we examined the performance of surface patterns that only contain part-of-speech information about target candidate pairs, such as [*the difference between <ANT/Adj> and <ANT/Adj>*]. In the third study, we examined the performance of automatically generated dependency patterns. Such patterns

contain syntactic dependencies and abstract away from the surface structure, so they can identify opposites that co-occur with each other within a sentence too far away.

The results show that all three pattern-based methods can find good opposites. However, the methods differ in the performance as to the number and range of opposites they can identify. In comparison to other methods, patterns with part-of-speech information find the largest number of opposites that include not only already known pairs but also novel opposites that are usually not studied or discussed by theoretical linguists, and opposites that are contrastive only in certain contexts and domains. Opposites found with part-of-speech patterns would be useful for many computational applications, including automatic identification of Contrast.

Strictly textual patterns also find good opposites but they find fewer pairs overall and the majority of pairs they find are well-known opposites. Because there are no restrictions as to the syntactic categories of found pairs, the most productive textual patterns tend to find the same pairs in the top results with the seeds expressed by adjectives and nouns and to a lesser degree with the seeds expressed by verbs.

Dependency patterns also find well-established and novel pairs of opposites but fewer than part-of-speech patterns, especially at finding opposites expressed by verbs.

Taking all this into consideration, we conclude that the best method for finding opposites automatically is an algorithm that uses part-of-speech patterns.

Our results have several implications for the research on antonymy. The first implication concerns the ongoing discussion as to which pairs can be treated as antonymous. While the concept of antonymy has been mostly discussed in relation to opposites expressed by adjectives, all three pattern-based methods found the largest number of opposites with the seeds expressed by nouns. Nominal opposites are tricky to categorize, as it is often unclear whether a noun - noun pair should be treated as regular co-hyponyms or as opposites. This is where the context, in which a pair is found, can be used to determine whether a pair is contrastive or not. Especially because we do not extract any kinds of co-hyponyms, but only those pairs that are contrastive in nature. For example, while our algorithms found the pair *cat - dog*, they did not find pairs with other members of the category ANIMALS, like *cat - horse*. Overall, given that nominal opposites are so persistent in the results, there is a need for a classification of opposites that will cover the variety of pairs we find automatically.

The second implication concerns the topic of antonym canonicity, that is, an intuitive difference between canonical or "good" opposites like *fast - slow*, *old - young* and non-canonical "less good" opposites like *slow - rapid* and *young - aged*. In the

past, it has been suggested that the difference in canonicity can be explained by the fact that canonical opposites co-occur with each other in a larger number of pattern types. However, our results from a corpus-driven approach show that both canonical and non-canonical pairs are likely to occur in the same types of patterns equally often. Therefore, it is necessary to study further the connection between patterns, their types and what they can say about antonymy.

The work presented in this dissertation is a promising first step towards a better understanding of opposites, their behaviour and functions in discourse. The corpus-driven aspect of this approach is crucial as it is unbiased towards researchers‘ intuitions and provides an objective way of studying the fascinating world of opposites. In the future, it would be interesting to test whether the results of the pattern-based methods can be automatically evaluated using distributional methods that have been successful at validating existing opposites rather than finding novel pairs. Further, it would be interesting to extend the concept of opposites to cross-categorical pairs like *begin - endless* (verb - adjective), which also indicate some type of contrast, and test whether pattern-based methods can find such pairs.

Publications

Lobanova, A., van der Kleij, T. and J. Spenader (2010). Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: *International Journal of Lexicography*. Vol 23: 19-53.

Lobanova, A. (2010). The role of prominence scales on disambiguation of grammatical functions in Russian. In: *Russian Linguistics*. Vol 35(1).

Lobanova, A., Bouma, G. and Tjong Kim Sang, E. (2010). Using a treebank for finding opposites. In: *Proceedings of TLT9*, Tartu, Estonia, pp.139-150.

Lobanova, A. (2009). Expressiveness of 'a'. In: *International Journal for Language Data Processing (SDV)*. 1-2: 95-108.

Lobanova, A., Spenader, J, van de Cruys, T., van der Kleij, T. and Tjong Kim Sang E. (2009). Automatic relation extraction - can synonym extraction benefit from antonym knowledge? In: *NODALIDA2009 workshop WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense, Denmark.

Spenader, J. and Lobanova A. (2009). Reliable discourse markers for Contrast. In: *Proceedings of the Eighth International Workshop on Computational Semantics*, Tilburg, the Netherlands.

Lobanova, A. (2007). Versatility and restrictions on the use of Russian 'a'. In: *Proceedings of the Fifth Semantics in the Netherlands Day*, pp.13-26, 28 September, Groningen, the Netherlands.

Lobanova, A. and Spenader J. (2007). Incorporating polarity in lexical resources. In: *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon*, 10-11 May, Paris, France.

Lobanova, A., Spenader, J., and Valkenier B. (2007). Lexical and perceptual grounding of a sound ontology. In: *Matousek, V. and P. Mautner (Eds.): Text, Speech and*

Dialogue, 10th International Conference, pp.180-187, 3-7 September, Pilsen, Czech Republic.

Bibliography

Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, 1999. 24, 37, 41, 42, 47, 50, 51, 176

Gosse Bouma, Jori Mur, and Gertjan van Noord. Reasoning over dependency relations for QA. In *In Knowledge and Reasoning for Answering Questions (KRAQ05), IJCAI Workshop*, pages 15–21, 2005. 177

Thorsten Brants and Alex Franz. *Web 1t 5-gram version 1*. Linguistics Data Consortium, 2006. 29

Lou Burnard. *Reference guide for the British National Corpus*. Oxford University Computing Services, 2000. 30, 48, 49

Walter G. Charles and George A. Miller. Contexts of antonymous adjectives. *Applied Psycholinguistics*, 10(3):357–375, 1989. 19, 30, 48, 61, 63, 102, 184, 240, 246

Eugene Charniak. A maximum-entropy-inspired parser. In Janyce Wiebe, editor, *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139, Seattle, Washington, 2000. Morgan Kaufmann Publishers, San Francisco, CA, USA. 181

Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136–143. Austin, Texas, 1988. 58

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 26, 29, 178, 181

Michael Collins. A new statistical parser based on bigram lexical dependencies. In *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-1996)*, pages 184–191, 1996. 181

Alan Cruse. *Lexical semantics*. Cambridge: Cambridge University Press, 1986. 9, 11, 12, 13, 14, 16, 54, 86, 225

Marie-Catherine de Marneffe, Anna Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, 2008. 3, 4

James E. Deese. The associative structure of some common English adjectives. *Journal of Verbal Learning and verbal Behaviour*, 3(5):347–357, 1964. 19, 48, 54

James E. Deese. *The structure of associations in language and thought*. The Johns Hopkins Press, 1965. 19

Christiane Fellbaum. Co-occurrence and antonymy. *International Journal of Lexicography*, 8:281–303, 1995. 20, 21, 23, 33, 49, 52, 77, 99, 113, 114, 167, 216, 240, 246

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998. 3, 24, 25, 29, 35, 36, 38, 63, 177

Joseph Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. 43

Gregory Grefenstette. Finding semantic similarity in raw text: the Deese antonyms. 1992. 28, 114

Michael A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. London: Longman, 1976. 33

Zellig S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954. 27

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545, 1992. 24, 25, 36, 38, 50, 51, 110, 111, 124, 177, 178

Katja Hofmann and Erik Tjong Kim Sang. Automatic extension of non-English WordNets. In *Proceedings of SIGIR-2007*. Amsterdam, The Netherlands, 2007. 38, 39, 50

Ales Horak, Piek Vossen, and Adam Rambousek. The development of a complex-structured lexicon based on WordNet. In *Proceedings of the 4th International GlobalWordNet Conference (GWC-2008)*, pages 200–208. Szeged, Hungary, 2008. 35, 38

Ashwin Ittoo and Gosse Bouma. On learning subtypes of the part-whole relation: do not mix your seeds. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 1328–1336, Uppsala, Sweden, 2010. 41, 47, 50, 178, 204, 212

Howard Jackson. *Words and their meaning*. London: Longman, 1988. 17

Valentin Jijkoun, Maarten de Rijke, and Jori Mur. Information extraction for Question Answering: improving recall through syntactic patterns. In *In Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 1284–1290, 2004. 177

Steven Jones. *Antonymy: a corpus-based perspective*. London: Routlegde, 2002. 11, 13, 15, 21, 22, 23, 47, 49, 61, 72, 83, 90, 100, 102, 112, 115, 118, 119, 132, 133, 164, 173, 174, 194, 211, 216, 217, 218, 240, 246

Steven Jones, Carita Paradis, Lynn Murphy, and Caroline Willners. Googling for opposites - a web-based study of antonym canonicity. *Corpora*, 2(2):129–155, 2007. 21, 22, 23, 28, 47, 49, 52, 54, 80, 86, 92, 100, 113, 118, 119, 133, 139, 168, 217, 219, 220, 240, 246

John S. Justeson and Slava M. Katz. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17:1–19, 1991. 19, 23, 30, 33, 48, 173, 240, 246

Maire Weir Kay, editor. *Webster's Collegiate Thesaurus*. Merrian-Webster, 1988. 32

Ruth M. Kempson. *Semantic theory*. Cambridge: Cambridge University Press, 1977. 13

Henry Kučera and Jindřich Francis, editors. *Computational analysis of presentday American English*. Providence, RI: Brown University Press, 1967. 5, 19

Richard Landis and Gary Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. 45, 241, 247

Adrienne Lehrer and Keith Lehrer. Antonymy. *Linguistics and Philosophy*, 5:483–501, 1982. 12

Dekang Lin and Patrick Pantel. Discovery of inference rules for Question Answering. *Natural Language Engineering*, 7:343–360, 2001. 177, 181

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI-2003*, pages 1492–1493, 2003. 4, 31, 32, 133

Anna Lobanova, Jennifer Spenader, Tim van de Cruys, Tom van der Kleij, and Erik Tjong Kim Sang. Automatic relation extraction - can synonym extraction benefit from antonym knowledge? In *Proceedings of NODALIDA 2009 Workshop 'WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies'*, 2009. 4

Anna Lobanova, Tom van der Kleij, and Jennifer Spenader. Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography*, 23(1):19–53, 2010. 218, 221, 224

John Lyons. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press, 1968. 11

John Lyons. *Semantics*, volume 1. Cambridge: Cambridge University Press, 1977. 11, 12, 13, 16, 79, 166

Isa Maks, Willy Martin, and H. de Meerseman, editors. *Referentie Bestand Nederlands. Manual*. Vrije Universiteit, Amsterdam, 1999. 38

Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, 2002. 3, 100, 114

W. Martin and G. A. J. Tops. *Groot woordenboek Engels-Nederlands*. Van Dale Lexicografie. Utrecht, 1986. 40

W. Martin and G. A. J. Tops. *Groot woordenboek Nederlands-Engels*. Van Dale Lexicografie. Utrecht, 1989. 40

Paul McNamee, Rion Snow, Patrick Schone, and James Mayfield. Learning named entity hyponyms for Question Answering. In *In Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 799–804, 2008. 178

Rada Mihalcea and Carlo Strapparava. Making computers laugh: investigations in automatic humor recognition. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005. 4

Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: using Mechanical Turks to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010. 4, 222

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. Computing word-pair antonymy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 982–991, 2008. 4, 28, 30, 31, 33, 116

M. Lynne Murphy. *Semantic relations and the lexicon: antonyms, synonyms and other semantic paradigms*. Cambridge: Cambridge University Press, 2003. 11, 18, 145, 167, 226

Roeland J. F. Ordelman, editor. *Twente Nieuws Corpus (TwNC)*. Parlevink Language Technology Group. University of Twente, 2002. 54, 113, 128, 180

Frank Palmer. *Semantics: a new outline*. Cambridge: Cambridge University Press, 1976. 17

Patrick Pantel and Marco Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113–120, 2006. 26, 50, 86, 98, 106, 148, 178, 179, 181, 182, 193, 204, 212

Carita Paradis and Caroline Willners. Antonyms in dictionary entries: methodological aspects. *Studia Linguistica*, 61(3):261–277, 2007. 22, 41, 49

Carita Paradis, Caroline Willners, and Steven Jones. Good and bad opposites: using textual and experimental techniques to measure antonym canonicity. *The Mental Lexicon*, 4(3):380–429, 2009. 49, 223

Justus J. Randolph. Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa. 182

Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Aritficial Intelligence*, pages 1044–1049. The AAAI Press/MIT Press, 1996. 59

John Sinclair, editor. *Collins Cobuild advanced learner's English dictionary*. 2003. 36

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In Yair Weiss Lawrence K. Saul and Leon Bottou, editors, *In Proceedings of the NIPS 17*, pages 1297–1304. MIT Press, 2005. 38, 39, 50, 124, 171, 172, 176, 177, 178, 179, 240, 247

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 801–808, 2006. 38, 50, 172

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? Evaluating non-expert annotations for Natural Language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 254–263. Association for Computational Linguistics, 2008. 171

Jennifer Spenader and Gert Stulp. Antonymy in Contrast relations. In *Seventh International Workshop on Computational Semantics*, 2007. 3, 100

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL-2008*. Manchester, UK, 2008. 176

Michael Thelen and Ellen Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, pages 214–221, 2002. 59

Erik Tjong Kim Sang and Katja Hofmann. Automatic extraction of Dutch hypernym-hyponym pairs. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 163–174, 2007. 50, 124, 127, 172, 176, 177

Erik Tjong Kim Sang and Katja Hofmann. Lexical patterns or dependency patterns: which is better for hypernym extraction? In *Proceedings of CoNLL-2009*, pages 174–182. Boulder, CO, USA, 2009. 50, 119, 123, 124, 125, 126, 127, 172, 176, 179, 202, 212, 217, 240, 247

Peter D. Turney. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 905–912. Manchester, UK, 2008. 32

Lonneke van der Plas and Gosse Bouma. Syntactic contexts for finding semantically similar words. In *Proceedings of the Computational Linguistics in the Netherlands (CLIN)*, pages 173–186, 2005. 39, 176, 241, 247

Lonneke van der Plas and Jörg Tiedemann. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, 2006. 39

Gertjan van Noord. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *In TALN 2006. Verbum Ex Machina. Actes de la 13e Conference sur le Traitement Automatique des Langues Naturelles*, pages 20–42, 2006. 54, 125, 127, 174, 181

Ellen M. Voorhees. Contradictions and justifications: extensions to the textual entailment task. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, pages 63?–71, 2008. 3

Piek Vossen, Laura Bloksma, and Paul Boersma. The Dutch WordNet. In *EuroWordNet Paper*. University of Amsterdam, 1999. 38

Caroline Willners and Carita Paradis. Swedish opposites: a multi-method approach to 'goodness of antonymy'. In *Lexical-Semantic Relations: Theoretical and Practical Perspectives)*, pages 15–47, 2010. 2, 81, 86, 217

English Summary

Chapter 1: Introduction

This dissertation deals with opposites, that is, words like *rich - poor*, *day - night*, *open - close*, and other pairs that express some type of contrast. In particular, we address two research questions. First, we explore pattern-based methods for finding opposites automatically. Pattern-based methods are commonly used to automatically extract meronyms (*car - wheel*) and hyponyms (*dog - animal*). However, no work has been done in this direction on finding opposites, although automatically found opposites would be useful for many natural language processing tasks, including identification of the *discourse relation of contrast* and identification of irony and contradictions. Second, we analyze automatically found opposites and compare them with opposites extensively studied and classified by theoretical linguists.

Opposites are easier to illustrate than to define and, as a result, many different classifications have been proposed in the past. This led to more confusion than consistency, especially with opposites expressed by syntactic categories other than adjectives. In contrast, a corpus-driven approach taken in this work provides methodologically-sound and objective means to studying opposites from real data usage.

Chapters 2 and 3: Theoretical framework, evaluation methodology and predictions

Chapter 2 introduces in detail corpus-based studies on opposites that laid the foundation for the current work. In particular, our three pattern-based methods are based on the assumption that opposites co-occur with each other within a sentence significantly more often than would be expected by chance and that often they can be found in intrasentential patters like [*between <ANT> and <ANT>*]. This has been known as the *Co-occurrence Hypothesis*, originally proposed by Charles and Miller [1989] and further developed and studied by Justeson and Katz [1991], Fellbaum [1995] and more recently by Jones [2002].

The hypothesis is valid not only for opposites expressed by adjectives but also by nouns and verbs, although the latter tend to be found in patterns in larger corpora (Fellbaum [1995], Jones [2002]). Based on this, we assume that our algorithms can successfully find patterns on the sentence level that can find good opposites expressed by adjectives, nouns and verbs, although the size of the corpus will play a role in that more nominal and verbal opposites will be found in corpora of larger size. Further, we can use significant co-occurrence to eliminate noise from the results.

Based on the previous work that aims to explain the intuitive differences between typical "good" opposites like *rich - poor*, *short - long*, *young - old* and less-typical opposites like *classical - popular*, *green - grey*, by arguing that the former co-occur with each other in more pattern types than the latter and, as a result, they are stronger associated with each other (Jones et al. [2007]), we assume to find typical opposites in a wider range of automatically generated pattern types.

An important contribution of our work is the study of the types of patterns that perform best. In particular, first we test *surface patterns* that do not contain any syntactic information. Then, we test surface patterns that only contain part-of-speech information about the target pairs, so-called *part-of-speech patterns*. Finally, we use a fully parsed corpus to generate *dependency patterns* that contain syntactic dependencies. There is no consensus in the computational community as to how much of syntactic information is needed for the best results, with some researchers showing that the best results for hyponym - hypernym extraction are achieved with dependency patterns (Snow et al. [2005]) and others - with patterns that contain part-of-speech information only (Tjong Kim Sang and Hofmann [2009]). We assume that patterns with syntactic information are especially beneficial for finding opposites expressed by nouns and

verbs.

Finally, we compare the results for two corpora of different genre, namely, newspaper texts and encyclopedia texts. We assume that the genre of the corpus plays a role as encyclopedia texts, often used for automatic extraction of meronyms and hyponyms, exhibit repetitive structures and, unlike newspaper texts, do not provide enough variation for identification of various pattern types.

Chapter 2 also provides a thorough description of the existing theoretical approaches to opposites, showing their limitations and implications for the current work. Although researchers have proposed many classifications, we show that none of them provide reliable means of separating opposites from non-opposites, especially when dealing with non-conventional pairs. As a result, it is difficult to evaluate found candidate pairs, as many of them are novel and do not fall under any of the previously proposed classes of opposites. Because of this, we use several methods for the evaluation of found pairs, all discussed in detail in Chapter 3.

First, we use two existing lexical resources in Dutch, namely, CORNETTO and *Mijnwoordenboek.nl*, to evaluate found pairs. This evaluation method is often used in the work on automatic relation extraction, however, as has been previously shown (van der Plas and Bouma [2005]), such resources often miss good pairs and as a result they are not sufficient for the evaluation of the results. Manual evaluation is the second common way to evaluate results in the work on automatic relation extraction. So, we also asked three participants to evaluate found pairs with the scoring above a given threshold. To ensure that such evaluation is reliable we calculate inter-annotator agreement, following the scale originally proposed by Landis and Koch [1977]. Finally, based on the classification by the participants, the precision scores are calculated in order to compare our results to similar work on meronym and hyponym extraction.

Chapters 4 - 6: Experiments and results

Using small sets of six, 12 and 18 seed pairs expressed either by adjectives, nouns or verbs, we identify the best patterns for finding new pairs of opposites in a 450 million word newspaper corpus of Dutch. In the first study, discussed in Chapter 4, we automatically generate strictly textual patterns like [*either <ANT> countries or <ANT> countries*] that do not contain any syntactic information, but simply capture surface strings. In the second study, presented in Chapter 5, we generate surface patterns

that contain part-of-speech information about target word pairs, like [*the difference between <ANT/Adj> and <ANT/Adj>*]. In the third study, presented in Chapter 6, we use a parsed corpus to automatically acquire patterns with syntactic dependencies. Such patterns abstract away from the surface structure capturing that, for example, *<ANT1/Noun>* is the subject and *<ANT2/Noun>* is the direct object and they are connected by the verb *appreciate*.

The best results were achieved with part-of-speech patterns (Chapter 5), which identified many typical as well as novel opposites. For example, with the set of 18 adjective - adjective seeds, part-of-speech patterns found 517 pairs that were judged as opposites by at least two participants, leading to the precision of 0.6 for the top-100 pairs. In comparison, using the same seed set, textual patterns found 208 pairs and dependency patterns found 169 pairs that were judged as opposites by at least two participants.

The same tendency was found for the results with noun - noun and verb - verb seed sets, although verbs extracted the least number of opposites with all pattern types. In particular, using the set of 18 noun - noun seeds, part-of-speech patterns extracted 399 pairs that were judged as opposites by at least two participants, leading to the precision of 0.61 for the top-100 pairs. In comparison, textual patterns found 220 opposites and dependency patterns found 141 opposites, according to the participants. Using the set of 18 verb - verb seeds, part-of-speech patterns found 87 pairs judged as opposites by the participants, leading to the precision score of 0.56 for the top-100 pairs. With the same seed set, textual patterns found 43 opposites and dependency patterns found 78 pairs judged as opposites by the participants. Thus, given the same seed set and the same corpus, part-of-speech patterns identified the largest number of opposites across all three syntactic categories.

The main limitation of textual patterns (Chapter 4) is that they find the same most frequent opposites across the seed sets of all three syntactic categories and the majority of these pairs are well-established opposites. This means that textual patterns are useful only as a simple method that requires no preprocessing of the corpus and that can identify the most common opposites across different syntactic categories with high precision.

Although dependency patterns (Chapter 6) found the least number of opposites per seed set according to the participants and lexical resources, similar to part-of-speech patterns, they found many novel pairs. Interestingly, dependency patterns and part-of-speech patterns found different kinds of opposites expressed by nouns among the

top found pairs. In particular, among top opposites found with part-of-speech patterns, many were opposites of kinship like *grandmother - grandfather*, *brother - sister* and relational opposites like *teacher - student*. Opposites found with dependency patterns often expressed contrast between abstract nouns like *science - religion*, *strength - weakness*. This points out that the most productive part-of-speech patterns were qualitatively different from the most productive dependency patterns.

Overall, the best results are achieved by the algorithm that relies on adding the minimum amount of syntactic information, namely only part-of-speech information. Since this method does not require any computationally costly preprocessing steps and can easily be applied to vast amounts of data, part-of-speech patterns offer a promising solution to automatic extraction of opposites.

In Chapter 4 we also looked at the role of the genre of the corpus, comparing the results from the algorithm with strictly textual patterns run on the corpus of newspaper texts and a collection of encyclopedia texts. Our results suggest that the genre of the corpus matters and that the newspaper corpus yields much better results than the collection of encyclopedia texts. This is due to a more varied structure of sentences in the newspaper texts and, as a result, a wider range and number of productive patterns. In relation to the studies on relation extraction, in particular, meronyms, this means that there can be differences as to the most productive patterns and it might be that while encyclopedia texts provide a smaller number of frequent reliable pattern types, the newspaper texts contain more varied, less typical patterns. How this can affect the results needs to be tested in the future.

Chapters 7: Conclusions

In the final chapter the results of the experiments are discussed in relation to the questions raised in Chapter 2. The results show that the range of automatically found opposites surpasses the limited number of well-established opposites commonly discussed in the theoretical approaches on opposites. In particular, pattern-based methods can find not only typical opposites like *old - new*, *rich - poor*, but also less conventional opposites like *new - existing*, *new - second-hand*, *new - known* and *old - recent*, non-typical domain-specific opposites like *white - red* (wine), *Democrat - Republican* (political parties) and context-dependent pairs like *migrant - Dutchman* (Dutch newspaper texts), *foreign - Dutch* (as an analogue of *foreign - domestic* in the context of *local* and

international policies). Although such pairs exhibit similar behaviour in the corpus to the canonical opposites, non-typical context-dependent opposites have been neglected in theoretical classifications. Our results provide evidence that opposites include a much wider range of pairs than has been previously recognized.

In fact, automatically found opposites, especially domain-specific and context-dependent pairs that are often missed in the existing lexical resources, are particularly useful for other natural language processing tasks. This is further confirmed by the fact that, contrary to our assumptions, we found no differences between typical and non-typical opposites as to the frequency and the types of patterns in which they were found. This shows that both types are valid opposites that need to be studied in the future.

At the moment, the evaluation of the results is constrained by the fact that many good opposites are missing in the existing lexical resources and it is difficult to train participants to classify pairs as our algorithms found many non-typical opposites. In the future, it would be interesting to test whether the results of the pattern-based methods can be automatically evaluated using distributional methods that have been successful at validating existing opposites rather than finding novel pairs.

Further, it would be interesting to extend the concept of opposites to cross-categorical pairs like *ask - answer* (verb - noun), *midland - foreign* (noun - adjective) that also indicate some type of contrast, and test whether pattern-based methods can find such pairs.

In short, the work presented in this dissertation is a promising first step towards a better understanding of opposites, their behaviour and functions in *discourse*. The corpus-driven aspect of this approach is crucial as it is unbiased towards researchers' intuitions and provides an objective way of studying the fascinating world of opposites.

Nederlandse Samenvatting

Hoofdstuk 1: Introductie

Dit proefschrift behandelt antoniemen: woordparen zoals *arm - rijk*, *dag - nacht*, *open - sluiten*, en andere paren die onderling een contrastrelatie uitdrukken. In het bijzonder richten we ons op twee onderzoeksvragen. Ten eerste bestuderen we patroon-gebaseerde methoden om automatisch antoniemen te vinden. Patroon-gebaseerde methoden worden vaak gebruikt om meroniemen (*auto - stuur*) en hyponiemen (*hond - dier*) automatisch te identificeren, maar voor antoniemen is deze methode nog niet eerder toegepast. Dit ondanks het feit dat het automatisch identificeren van antoniemen nuttig zou kunnen zijn voor veel toepassingen van natuurlijke-taalverwerking, zoals het herkennen van de *rhetorische contrastrelatie* en het herkennen van ironie en contradicties. Ten tweede analyseren we automatisch gevonden antoniemen en vergelijken we die met antoniemen die door theoretisch taalkundigen uitgebreid onderzocht en geclassificeerd zijn.

Antoniemen zijn makkelijker te illustreren dan te definiëren, en als gevolg daarvan zijn er al veel verschillende classificaties bedacht. Dit heeft geleid tot meer chaos dan consistentie, in het bijzonder met betrekking tot andere syntactische categorieën dan bijvoeglijke naamwoorden. Daarentegen biedt de corpusgebaseerde aanpak die wij in dit proefschrift hanteren een methodologisch verantwoorde en objectieve manier om tegenstellingen te bestuderen door echte data te gebruiken.

Hoofdstuk 2 en 3: Theoretisch kader, evaluatiemethodologie en voorspellingen

Hoofdstuk 2 introduceert in detail de corpusgebaseerde studies die de basis hebben gelegd voor ons huidige onderzoek. In het bijzonder zijn onze drie patroon-gebaseerde methoden gestoeld op de aanname dat antoniemen significant vaker samen voorkomen binnen zinnen dan anders verwacht zou worden, en verder dat ze gevonden kunnen worden in specifieke binnenzinse patronen, zoals [*tussen* <*ANT*> *en* <*ANT*>]. Dit staat bekend als de 'Co-occurrence Hypothese', oorspronkelijk voorgesteld door Charles en Miller Charles and Miller [1989] en verder ontwikkeld en bestudeerd door Justeson en Katz Justeson and Katz [1991], Fellbaum Fellbaum [1995], en recenter, door Jones Jones [2002].

De hypothese geldt niet alleen voor antoniemen die worden uitgedrukt door bijvoeglijke naamwoorden, maar ook voor zelfstandige naamwoorden en werkwoorden, hoewel de laatste vooral gevonden worden in patronen in grotere corpora (Fellbaum [1995], Jones [2002]). Om deze reden nemen we aan dat onze algoritmes in staat zijn om patronen te vinden op zinsniveau, die het mogelijk maken om tegenstellingen te identificeren die bestaan uit paren bijvoeglijke naamwoorden, zelfstandige naamwoorden en werkwoorden. Hierbij verwachten we dat de grootte van de corpus een rol zal spelen, en dat in grotere corpora relatief meer tegenstellingen bestaande uit zelfstandige naamwoorden en werkwoorden gevonden zullen worden. Verder kunnen we ruis uit onze resultaten verwijderen door te kijken welke paren significant vaker dan verwacht samen voorkomen.

Eerder werk verklaart de intuïtieve verschillen tussen typische "sterke" antoniemen (zogenoemde '*canonical opposites*') als *arm - rijk*, *kort - lang*, *jong - oud* en minder typische tegenstellingen als *klassiek - populair* and *groen - grijs* door ervan uit te gaan dat de eerstgenoemde in meer verschillende typen patronen samen voorkomen, en dat ze hierdoor sterker met elkaar geassocieerd zijn (Jones et al. [2007]). Op basis hiervan nemen we aan dat wij dergelijke typische antoniemen zullen vinden in meer verschillende automatisch gegeneerde patroontypen.

Een belangrijke bijdrage van ons werk is het bestuderen van de typen patronen die het beste presteren. In het bijzonder testen we eerst oppervlakte patronen (zogenoemde '*surface patterns*') die geen syntactische informatie bevatten. Daarna testen we patronen die alleen woordsoortinformatie bevatten over de doelparen (zogenoemde '*part-of-speech patterns*'). Tenslotte gebruiken we een volledig syntactisch (taalkundig) geanalyseerd corpus om patronen te genereren die syntactische afhankelijkheden bevat-

ten (zogenoemde '*dependency patterns*'). Er bestaat in de computationele-taalkunde-gemeenschap geen consensus over hoeveel syntactische informatie noodzakelijk is voor het behalen van de beste resultaten. Sommige onderzoekers laten zien dat de beste resultaten voor hyponiem-hyperniemextractie worden behaald met patronen die gebruik maken van syntactische afhankelijkheden (Snow et al. [2005]), terwijl anderen laten zien dat patronen die alleen woordsoortinformatie bevatten het beste werken (Tjong Kim Sang and Hofmann [2009]). Wij nemen aan dat patronen met syntactische informatie vooral nuttig zullen zijn voor het identificeren van antoniemen die worden uitgedrukt door zelfstandige naamwoorden en werkwoorden.

Tenslotte vergelijken we de resultaten van twee verschillende typen corpora, namelijk een collectie krantenteksten en een collectie encyclopedieteksten. Wij nemen aan dat het type corpus een rol speelt, aangezien encyclopedieteksten vaak herhalende structuren gebruiken. Zij worden vaak gebruikt voor het automatisch identificeren van meroniemen en hyponiemen, maar in tegenstelling tot krantenteksten bevatten zij niet genoeg variatie voor het identificeren van verschillende patroontypen.

Hoofdstuk 2 bevat ook een uitgebreide beschrijving van de bestaande theoretische visies op tegengestelde woorden, waarbij hun beperkingen en implicaties voor het huidige onderzoek worden besproken. Hoewel onderzoekers verschillende classificaties hebben voorgesteld, laten wij zien dat geen van die classificaties in staat is om op een consistente manier antoniemen van niet-antoniemen te onderscheiden, vooral waar het gaan om niet-conventionele paren. Als gevolg hiervan is het moeilijk om kandidaatparen te evalueren, aangezien vele daarvan nieuw zijn, en niet behoren tot een van de eerder voorgestelde typen antoniemen. Om deze reden gebruiken we verschillende manieren om de gevonden paren te evalueren, die allemaal in detail worden besproken in Hoofdstuk 3.

Ten eerste gebruiken we twee bestaande lexicale bronnen in het Nederlands, namelijk CORNETTO en *Mijnwoordenboek.nl*, om gevonden paren te evalueren. Deze evaluatiemethode wordt vaak gebruikt bij onderzoek naar automatische relatieherkenning. Er is echter aangetoond dat 'goede paren' vaak ontbreken in dergelijke bronnen (van der Plas and Bouma [2005]), en zodoende zijn zij niet voldoende voor het evalueren van onze resultaten. De op-één-na meest gebruikte evaluatie methode is handmatige beoordeling. Daarom hebben wij drie deelnemers gevraagd om alle gevonden paren met een voldoende hoge score te evalueren. Om de betrouwbaarheid van de evaluatie te verzekeren, hebben we de overeenkomst tussen de verschillende deelnemers berekend, volgens de schaal die oorspronkelijk is voorgesteld door Landis and Koch

[1977]. Tenslotte hebben we, op basis van de classificaties van de deelnemers, precisie scores berekend, om onze resultaten te kunnen vergelijken met soortgelijk werk over meroniem- en hyponiemextractie.

Hoofdstuk 4 - 6: Experimenten en resultaten

Met kleine verzamelingen van zes, 12 en 18 initiële 'kiemparen' (zogenoemde '*seeds*'), uitgedrukt in ofwel bijvoeglijke naamwoorden, ofwel zelfstandige naamwoorden, ofwel werkwoorden, identificeren we de beste patronen voor het vinden van antoniemen in een krantencorpus van 450 miljoen Nederlandse woorden. In het eerste onderzoek, besproken in Hoofdstuk 4, genereren we automatisch zuiver tekstuele patronen, zoals [*of* <*ANT*> *landen of* <*ANT*> *landen*], die verder geen syntactische informatie bevatten. In het tweede onderzoek, beschreven in Hoofdstuk 5, genereren we tekstuele patronen die woordsoortinformatie bevatten, zoals [*het verschil tussen* <*ANT/Adj*> *en* <*ANT/Adj*>]. In het derde onderzoek, omschreven in Hoofdstuk 6, gebruiken we een ontleed en gelabeld corpus om automatisch patronen met syntactische afhankelijkheden te vinden. Zulke syntactische-afhankelijkheidspatronen abstraheren weg van de tekstuele vorm van zinnen, en specificeren in plaats daarvan bijvoorbeeld dat <*ANT/Noun*> het onderwerp is en <*ANT/Noun*> het lijdend voorwerp en dat ze verbonden worden door het werkwoord *waarderen*.

De beste resultaten werden behaald met woordsoortpatronen (Hoofdstuk 5), die vele antoniemen identificeerden, zowel nieuw als conventioneel. Toen er bijvoorbeeld werd gezocht met 18 initiële 'kiemen' van de vorm bijvoeglijk naamwoord - bijvoeglijk naamwoord, werden er 517 paren gevonden die door tenminste twee deelnemers werden herkend als antoniemen. Hiermee was de precisie 0.6 voor de top-100 paren. Ter vergelijking, met dezelfde kiemverzameling werden er met zuiver tekstuele patronen 208 paren gevonden die door tenminste twee deelnemers werden herkend, en met syntactische-afhankelijkheidspatronen 169 paren.

De resultaten van 'kiemen' die bestonden uit zelfstandige naamwoorden en werkwoorden wezen in dezelfde richting, hoewel werkwoorden in alle gevallen leidden tot de identificatie van de minste antoniemen. Met de verzameling van 18 zelfstandig naamwoord - zelfstandig naamwoord 'kiemen' werden er 399 paren geïdentificeerd met de woordsoortpatronen, leidend tot een precisie van 0.61 voor de top-100 paren. Tekstuele patronen, daarentegen, vonden 220 tegenstellingen, en syntactische-afhanke-

lijkheidspatronen 141, volgens de beoordelingen van de deelnemers. Met de verzameling van 18 werkwoord - werkwoord 'kiemen' vonden woordsoortpatronen 87 antoniemen die ook door de deelnemers werden herkend, met een precisie score van 0.56 voor de top-100 paren. Met dezelfde kiemverzameling vonden tekstuele patronen 43 antoniemen en syntactische-afhankelijkheidspatronen 78 antoniemen, weer volgens de beoordelingen van de deelnemers. We kunnen dus concluderen dat gegeven dezelfde combinatie van initiële kiemverzameling en corpus, woordsoortpatronen de meeste antoniemen identificeerden, in alle drie de syntactische categorieën.

De belangrijkste beperking van tekstuele patronen (Hoofdstuk 4) is dat ze dezelfde veel voorkomende antoniemen vinden met de kiemverzamelingen van alle drie de syntactische categorieën, en dat de meeste van deze paren alom bekende, conventionele antoniemen zijn. Dit betekent dat tekstuele patronen alleen bruikbaar zijn als eenvoudige methode die geen voorverwerking van de corpus vereist, waarbij met hoge precisie de meest gebruikelijke antoniemen worden gevonden, in verschillende syntactische categorieën.

Hoewel de syntactische-afhankelijkheidspatronen (Hoofdstuk 6) de minste antoniemen vonden per kiemverzameling, volgens zowel de deelnemers als de lexicale bronnen, vonden zij wel veel nieuwe paren, net als de woordsoortpatronen. Hierbij was het interessant dat de syntactische-afhankelijkheidspatronen en de woordsoortpatronen verschillende typen antoniemen hadden binnen hun top-100 paren. Met de woordsoortpatronen werden vele familiegerelateerde antoniemen gevonden, zoals *opa - oma*, *broer - zus* en relationele tegenstellingen, zoals *docent - student*, *leerling - leraar*. Antoniemen gevonden door syntactische-afhankelijkheidspatronen drukten vaak contrast uit tussen abstracte zelfstandige naamwoorden, zoals *religie - wetenschap*, *kracht - zwakte*. Dit illustreert dat de meest productieve woordsoortpatronen kwalitatief verschilden van de meest productieve syntactische-afhankelijkheidspatronen.

Over het algemeen zijn de beste resultaten behaald door het algoritme dat uitging van de minimale hoeveelheid toegevoegde syntactische informatie, namelijk woordsoortinformatie. Aangezien deze methode geen zware computationele voorverwerking vereist en gemakkelijk kan worden toegepast op grote hoeveelheden data, zijn woordsoortpatronen een veelbelovende manier om antoniemen automatisch te identificeren.

In Hoofdstuk 4 hebben we ook gekeken naar de rol van het type corpus, waarbij we de resultaten van het algoritme dat gebruik maakte van strikt tekstuele patronen hebben vergeleken tussen een kranten-en een encyclopediecorpus. Onze resultaten suggereren dat het type corpus uitmaakt, en dat een collectie krantenteksten veel

betere resultaten oplevert dan een collectie encyclopedieteksten. Dit komt door de meer gevarieerde zinsstructuur in de kranten teksten, waardoor er meer en meer verschillende productieve patronen zijn. Voor andere studies over relationele extractie, in het bijzonder met betrekking tot meroniemen, betekent dit dat er verschillen kunnen zijn in welke patronen het meest productief zijn. Het zou kunnen dat encyclopedieteksten een kleiner aantal, veel voorkomende, betrouwbare patroontypen bevatten, terwijl in krantenteksten gevarieerdere, minder typische patronen voorkomen. Hoe dit de resultaten kan beinvloeden moet in de toekomst nog verder onderzocht worden.

Hoofstuk 7: Conclusies

In het laatste hoofdstuk worden de resultaten van de experimenten besproken in relatie tot de vragen die gesteld zijn in Hoofdstuk 2. De resultaten laten zien dat de automatisch gevonden tegenstellingen meer gevarieerd zijn dan het beperkte aantal typische antoniemen dat meestal besproken wordt in theoretische verhandelingen over antoniemen. In het bijzonder vinden patroon-gebaseerde methodes niet alleen alom bekende voorbeelden als *oud - nieuw*, *arm - rijk*, maar ook minder conventionele antoniemen als *nieuw - bestaand*, *nieuw - tweedehands*, *nieuw - bekend*, en *oud - recent*, atypische domeinspecifieke antoniemen als *wit - rood* (wijn), *Democraat - Republikein* (politieke partijen) en contextafhankelijke paren zoals *migrant - Nederlander* (Nederlandse krantenteksten), *buitenlands - Nederlands* (analoog aan buitenlands - binnenlands in de context van *lokaal* en *internationaal* beleid). Hoewel zulke paren zich binnen de corpora hetzelfde gedragen als canonieke antoniemen, worden atypische contextafhankelijke paren zelden meegenomen in theoretische classificaties. Onze resultaten laten zien dat antoniemen veel meer verschillende paren omvatten dan eerder erkend is.

Verder is het zo dat automatisch gevonden antoniemen, in het bijzonder de domeinspecifieke en contextafhankelijke paren die vaak ontbreken in bestaande lexicale bronnen, erg nuttig zijn voor andere taken binnen de natuurlijke taalverwerking. Dit wordt bevestigd door het feit dat, in tegenstelling tot onze eerdere aannames, we geen verschillen vonden tussen typische en atypische antoniemen met betrekking tot het totale aantal en het aantal verschillende patronen waarin zij voorkwamen. Dit laat zien dat beide typen echte antoniemen zijn die in de toekomst verder bestudeerd moeten worden.

Op het moment wordt het evalueren van onze resultaten bemoeilijkt door het het feit dat veel goede antoniemen ontbreken in bestaande lexicale bronnen, en doordat het lastig is om deelnemers te trainen in het beoordelen van de gevonden paren, aangezien onze algoritmes veel atypische antoniemen vinden. In de toekomst zou het interessant zijn om te toetsen of de resultaten ook automatisch geëvalueerd kunnen worden, door gebruik te maken van gedistribueerde methoden die nu ingezet worden om bekende antoniemen te verifiëren, in plaats van nieuwe paren te identificeren.

Verder zou er onderzocht kunnen worden of het concept 'antoniemen' ook kan worden uitgebreid naar paren uit verschillende categorieën die een contrast uitdrukken, zoals *vragen - antwoord* (werkwoord - zelfstandig naamwoord), *binnenland - buitenlands* (zelfstandig naamwoord - bijvoeglijk naamwoord), en of patroon-gebaseerde methoden zulke paren ook kunnen vinden.

Samenvattend is het werk dat in dit proefschrift wordt gepresenteerd een veelbelovende eerste stap naar een beter begrip van antoniemen, hun gedrag en hun rol in het *discours*. De corpusgebaseerde aspecten van de gebruikte aanpak zijn cruciaal, aangezien ze niet beinvloed worden door de intuïties van onderzoekers, en zodoende een objectieve manier bieden om de fascinerende wereld van antoniemen te bestuderen.