Defining Antonymy: A Corpus-based Study of Opposites Found by Lexico-syntactic Patterns

Abstract

Using small sets of adjectival seed antonym pairs, we automatically find patterns where these pairs co-occur in a large corpus of Dutch, and then use these patterns to extract new antonym pairs. Evaluation of extracted pairs by five human judges showed that automatic scores correlate with human evaluation and that pattern based methods can be used to extract new antonym pairs. The majority of extracted pairs were noun-noun pairs, contrary to expectations based on previous research. Additionally, the method identifies a subgroup of co-hyponyms that frequently function antonymously, and together with more traditional antonyms makes up a wider class of incompatibles, suggesting that antonymy is a diverse relation that includes pairs of different types and categories that are not captured by any single linguistic theory. Comparison with Dutch WordNet and an online Dutch dictionary shows that only a handful of extracted pairs are currently listed in these existing resources, emphasizing the usefulness of the project.

1. Introduction

The pairs *rich–poor* and *wealthy–destitute* are generally recognized as good antonyms. On the other hand, the pair *athletic–chubby* could be opposites in the right context, but could also be perceived as co-hyponyms, because both terms describe physical builds. Most speakers would consider the noun-noun pair *man–woman* to be opposites, but when asked to list other noun-noun pairs may not be likely to spontaneously list pairs like *haves–have nots* and *inner city–suburbia.* This points to the main problem with antonymy – despite being one of the most frequent lexical relations, it is poorly understood and not-well defined. Our goal is therefore twofold. First show how a simple method for extraction of good opposites automatically using a large corpus of text can find new antonym pairs. The results can be used to improve the coverage of antonyms in existing lexical resources and dictionaries, which can further be used for natural language processing applications. Second, we discuss the implications our results have on the definition of antonymy and classification of its types not currently captured by any linguistic theory.

Existing electronic dictionaries and lexical databases like WordNet (Fellbaum 1998) or corpus-based dictionaries such as COBUILD (see Paradis and Willners 2007 for a detailed survey) currently document only a small set of known, well-studied antonyms. Their coverage of antonyms in general is often extremely limited compared to other lexical relations. For this reason, an automatic method for antonym extraction would be particularly useful. Creating a comprehensive list of antonyms is also useful for many natural language applications that need some type of language understanding. More specifically, antonymous concepts may indicate contrastive relationships (Marcu and Echihabi 2002; Spenader and Stulp 2007) so antonyms could be used to help identify them. Antonym noise is also a known problem for automatic techniques that attempt to identify synonyms, hyponyms and other lexical relationships. An extended list of antonyms could be used as a filter to improve the performance of these techniques.

We automatically extract and evaluate candidate antonyms from a very large newspaper corpus of Dutch by using lexico-syntactic patterns. Starting with sets of "seed" pairs of well-established antonyms, we extract patterns in which the pairs co-occur, use the best ones to find new antonym pairs and then repeat this process to find new patterns. Pairs are automatically scored according to the frequency with which they occur in high scoring patterns. In addition, we evaluate our results with respect to Dutch WordNet and an online dictionary of Dutch, as well as having five human judges who also evaluated the top scoring pairs. The results show that the method finds many opposites not currently recognized as such in existing lexical resources, including many noun-noun antonym pairs and co-hyponyms that function contrastively. We argue that together these groups make a useful class of word pairs that are frequently used to express opposition and have been neglected in the existing literature.

We also argue that this method provides empirical evidence to support classifying some of the less well-established classes of opposites in theoretical literature, namely mutual incompatibles and contrastive co-hyponyms, as true antonym subtypes.

The paper is organized as follows. In section 2, we outline previous research on antonymy. Here we introduce the traditional classes of antonyms and the two major hypotheses that have influenced much existing corpus-based work, the substitutability hypothesis and the co-occurrence hypothesis. We also give an overview of studies where pattern-based methods were used to automatically extract other lexical relations. Section 3 explains our pattern-based method. Here we also give details of the automatic and manual evaluation we performed. Our results are presented in Section 4 and in Section 5 we discuss how our results compare to previous work on automatic lexical extraction, what implications it has for theoretical classes of antonymy, and how it relates to earlier predictions about the possibility of using pattern-based methods for extracting antonyms. Further, we outline how the results could be applied to natural language tasks. Finally, we present a number of ideas for future research in Section 6.

2. Previous research

2.1. *Defining antonymy*

Antonym pairs all express opposite or incompatible meanings. An example is the pair *rich–poor*, where we recognize that a person cannot be both rich and poor in the same way at the same time. This is a semantic definition of antonymy. Theoretical research has focused on semantic or logically based classifications of antonyms. Among analyzed antonym classes, there is, for example, a subset termed 'opposites', which includes pairs like *dead–alive*, *married–unmarried*. These are semantic opposites that exhaust the scale they refer to in that it is impossible to be married and unmarried at the same time. (Leech 1974, Lyons 1977, Cruse 1986). True 'antonyms' like *happy–sad,* on the other hand, are not mutually exclusive (it is possible to be neither happy nor sad) and unlike opposites, they are gradable. The most disputed category is called 'multiple incompatibles' (Lyons 1977). It 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 an instance 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*. Standard work on lexical relations would treat such pairs as co-hyponyms, and doesn't acknowledge this potential oppositional meaning. This can be due to the fact that all examples of incompatibles are with nouns, whereas the most scrutinized group of antonyms is made up of adjectives. In general, this previous theoretical work identifies a number of types of antonyms among known antonym pairs, but the research does not try to characterize antonyms from non-antonyms or identify new antonyms on the basis of the identified features.

These different kinds of semantic opposition have been intensely studied in the theoretical literature, but psycholinguistic studies of antonyms suggest these distinctions do not play a role in the way antonyms are represented in the mental lexicon. Stimulus-responses for word association tests have shown that a subset of adjectival antonyms have features 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; 1965). This type of strong response suggests that antonymy plays a role in the representation of adjectives in the mental lexicon. The set of antonym pairs that display this type of response includes pairs found in both the traditional category of opposites and true antonyms, (though *not* multiple incompatibles) and are frequently referred to as 'canonical' antonyms in subsequent research.

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 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 antonyms from the same pair were mutually interchangeable, participants would have no preferences as to the choice of adjectives, and should fill in each equally as often. However, the results showed that in many contexts, only one of the adjectives was appropriate. This was taken as evidence against the substitutability hypothesis. Instead Charles and Miller 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.

Justeson and Katz (1991) tested the co-occurrence hypothesis by examining the frequencies of intersentential occurrences of adjectival antonyms in the Brown Corpus. They confirmed that a set of adjectival antonyms co-occurred together significantly more often than sets of random adjectives.[1] Moreover, in many sentences they co-occurred in specific lexico-syntactic patterns like *between X and Y* and in these patterns antonyms *could* be substituted for one another. This result led to the conclusion that while frequent co-occurrence may be a characteristic of canonical antonyms, it isn't sufficient to establish the relationship because many lexically related words co-occur together significantly more often than chance. Instead co-occurring in certain intersentential patterns like *adjective–conjunction–adjective*, or in parallel constructions where an antonym pair modifies two identical nouns, is necessary for establishing the strong lexical association found in psycholinguistic tests. Because Charles and Miller looked at the sentences with only one of the two antonyms, the contexts were not always interchangeable. Justeson and Katz, on the other hand, examined only those sentences where antonyms co-occurred together; in such cases they could be substituted for one another.

These two hypotheses have been the foundation of the corpus-based work on antonymy. 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, 1965). Fellbaum (1995) conducted the first large-scale corpus work that looked at a wider class of antonym pairs, including nouns and verbs. She looked at the co-occurrences of nominal and verbal antonyms in the Brown Corpus and found that antonyms in both groups co-occurred with each other in the same sentence significantly more often than chance. However, unlike adjectival antonyms, they did not co-occur in parallel constructions or specific lexico-syntactic patterns. In fact, intersententially 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. What this predicts for our study is that by using lexico-syntactic patterns to extract antonyms automatically, we will probably only extract adjectival antonyms, because Fellbaum's results suggest nominal and verbal antonymous pairs will seldom occur in patterns.

Fellbaum also looked at the intersentential 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, suggesting that antonyms do not have to belong to the same part of speech. This implies, however, that they cannot be substituted for one another and they are unlikely to occur in patterns. Even more importantly, these findings suggest that semantic opposition will be frequently expressed with 'antonymous concepts', and will not be restricted to word pairs from the same syntactic category. Because of this Fellbaum suggests 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. Therefore, identifying pairs of words that express antonymous concepts, regardless of their syntactic functions would probably be useful for many natural language processing applications including automatic recognition of contrast relations in discourse. Although Fellbaum's results suggest that a pattern-based method like ours will be unlikely to identify cross-categorical antonyms, automatic extraction of a large set of antonyms within categories by means of patterns is an obvious research step that has to be tested first.

2.2. *Antonym canonicity and functions in discourse*

Much of the recent corpus based research on antonyms has focused on canonical antonyms and their properties. However, the actual defining characteristics of canonicity, and exactly which pairs can be considered canonical and which non-canonical is not at all clear. Such studies only relate to the goal of automatically acquiring antonym pairs in a limited way. Canonical antonyms themselves are of little interest: the set of canonical antonyms is finite, and well studied. It could nevertheless be useful as the features found with canonicity are possibly present to a lesser degree in other antonym pairs.

Jones et al. (2007) suggested that besides significant co-occurrence, the number of different patterns antonyms occur in, or their breadth of co-occurrence, should be used to determine which antonyms are canonical. Fourteen variations of seven lexico-syntactic patterns from Jones' earlier (2002) work were selected and Google was used to find patterns with pre-selected antonyms filling the first or the second adjective slot: *dull and X alike,* and *X and dull alike*. The variation helped to establish how reciprocal the antonymous relationship was to a given adjective pair. The retrieved antonyms were then ranked according to the number of different patterns the pair occurred in from the fourteen possible. 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.'s (2007) 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 information is that many antonymous pairs will not be very frequent even in very frequent patterns. The question is whether such pairs can still be retrieved and evaluated automatically by means of lexico-syntactic 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 antonyms in context. His goal was to identify the different textual functions of antonyms 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 selected from the total set of 55,411 extracted sentences, a proportional sample of 3,000 sentences manually. This sentence set was then used to define and classify lexico-syntactic patterns in which antonyms co-occurred.

Jones distinguished eight textual functions of canonical antonyms, of which six were indicated by lexico-syntactic patterns. The largest textual function with reliable patterns was Coordinated Antonymy, making up 38.4% [2] of all 3,000 examples. This function was found with patterns like *both X and Y*, *X or Y*, *X as well as Y* and as a group are said to "signal inclusiveness or exhaustiveness of scale" (Jones, 2002:61). Distinguished Antonymy, in which there is an emphasis on the distinction between the two groups, was characterized by patterns like *the difference between X and Y* and *separating X and Y*. Other textual functions included Comparative Antonymy (*more right than wrong*), Transitional Antonymy (*from success to failure*), Negated Antonymy (*in success, not failure*) and Extreme Antonymy (*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. [3] 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 2002:45, ex 5a) is used to emphasize another opposition (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 used, Jones relied on a list of antonyms 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 the major limitation of a pattern-based method for automatic extraction of antonyms in that lexico-syntactic patterns restrict the identification of antonyms 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 antonyms freely occurred in several types of patterns, the patterns do not coincide with the traditional classification of antonyms. Jones himself (2002: chapter 9) notes that he did not find any relationship between the traditional categories of antonyms and their textual functions.

Although these corpus-based studies give insight into the functions of antonyms in discourse and their canonicity, their main shortcoming is that the pairs as well as patterns used were identified manually, making the proportion of types unreliable. It is therefore not at all clear if, 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.

Despite these limitations, what the studies of Justeson and Katz (1991), Fellbaum (1995), Jones (2002) and Jones et al. (2007) *do* suggest is that a pattern-based method is a plausible way to extract some antonyms automatically. Automatic extraction of patterns will also have the advantage of not being influenced by biases or intuitions of the researcher. It is superior to manual identification because it requires less time and provides flexibility within different genres and languages, 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 antonym extraction. However, pattern-based methods have been used for automatic extraction of other lexical relations including hyponymy and meronymy. These studies are presented in Section 2.3.

2.3. *Automatic extraction of lexical relations*

Pattern based extraction methods can be traced back to Hearst's (1992) work on hyponyms. Using five manually identified lexico-syntactic patterns like *such X as Y* she found phrases, e.g. 'such authors as Shakespeare' and used them to successfully extract the fact that *Shakespeare* is a kind of author. Using a large corpus of encyclopedia texts (8.6 million words) as input, Hearst found 153 candidate hyponym pairs.[4] These pairs were made up of 226 unique words, of which 106 of the words were already in WordNet. Of this set 61 of these 106 words had the hypernym relationship listed in WordNet, and 45 were missing, suggesting that the method could easily add useful relations missing in WordNet.

As future work, Hearst suggested using seed pairs for given target lexical relations in order to find reliable patterns automatically. Testing Hearst's suggested automatic method, Berland and Charniak (1999) automatically extract meronyms using a very large corpus of 100 million words. Seeds, or a set of selected meronym pairs, are used to find sentences with potential patterns that were then manually identified. These patterns were then enlisted to extract new pairs. They report an accuracy of 55% for the top 50 meronyms derived for six examples based on the majority vote of the evaluation of the pairs by five human judges. They also compared the top twenty parts of the word *car* with WordNet and found that their method both missed important parts (e.g. *engine* and *door*) as well as found many parts not listed in WordNet (e.g. *tailpipe, break, speedometer).* Like Hearst (1992), Berland and Charniak did not automate the pattern identification step. As a result, it is not clear whether and how manual selection of patterns affected the results. For example, some patterns could be reliable, containing the target relation in most of the sentences (in other words, these patterns would have high precision), but at the same time they could be useful for only a small number of pairs (in other words, they would have low recall).

Using a minimally supervised bootstrapping algorithm, Pantel and Pennacchiotti (2006) identified generic patterns automatically. They were then used to extract a range of semantic relations including meronymy and hyponymy. Also beginning with seed pairs, they extracted all sentences these pairs co-occurred in and then generalized substrings from the patterns where they occurred. Pattern reliability of each string was calculated by means of pointwise mutual information to determine a set of reliable patterns. These patterns were then used to retrieve new pairs from the Web. Extracted pairs were evaluated using the association score between a given pair and a highly reliable pattern. They showed that the method of using automatically extracted generic patterns had high precision but also high recall. For hyponymy and meronymy in a newswire corpus (approx. 6 million words) they obtained precision scores between 73% and 85% with their top algorithm when a random sample of 50 extracted pairs were judged by two human judges. They did not compare their results to any existing lexical resources.

These previous results show that we should not expect to achieve precision scores above 70-80%, and we actually don't even know how realistic this prediction is for the current project as automatic antonymy extraction may be much more difficult than hyponym or meronym extraction. The greater challenge instead is in deciding on a good definition of antonymy itself. Meronymy and hyponymy are well-defined lexical relations for which there is a clear consensus among researchers about their definitions.The pattern-based studies described above relied heavily on existing lexical resources like WordNet to identify patterns and evaluate the results. Antonymy, on the other hand, is very little understood and poorly represented in WordNet. In our work, the starting point is only a small set of canonical pairs. This also makes the evaluation of our results more difficult. That is why in this study we not only aim to extract a large number of new antonym pairs, but we also hope that pairs extracted automatically can shed more light on what antonyms are, providing a wide range of different antonyms that fall outside the category of canonical. And pattern-based studies have shown quite clearly that automatic lexical extraction can be very fruitful, and more research might improve the state of the art.

3. Method

3.1. *Materials*

Antonym pairs were searched for using a normalized version of the CLEF (Cross-Language Evaluation Forum) corpus for Dutch (Jijkoun, Mishne, and de Rijke 2003). This corpus is made up of 72 million words and over four million sentences taken from the full content of the 1994 and 1995 Dutch daily newspapers *Algemeen Dagblad* and *NRC Handelsblad*.

We compiled two sets of canonical adjectival seeds: a set with six antonym pairs and a set with 18 canonical pairs where the first six were the same as in the six seed set. Seed sets are listed in Table 1. We used antonyms whose canonicity was established in earlier studies (Deese 1964, Jones et al. 2007) by word association tests and corpus analysis (breadth of co-occurrence) in order to maximize our chance of obtaining good patterns. We only use adjectives as seeds because these have been suggested to be the most frequently co-occurring antonyms in lexico-syntactic patterns (Jones 2002:110).

INSERT Table 1

3.2. *Algorithm*

The scope of this study is to test on the example of Dutch whether automatically acquired lexico-syntactic patterns can be used to extract antonyms. Since the role of the pre-processing of texts was not taken into account, the corpus was not lemmatized (that is words were not reduced to their basic forms – lemmas).

Our algorithm can be summarized in the following steps.

Step One: Extract all sentences that contain both halves of an antonym pair in the set of known (seed) antonyms.

Step Two: Replace all antonyms in the extracted sentences by some arbitrary wildcard token (for example, ‘<ant>’). Find the most frequently occurring patterns. (We used the approach outlined in Ravichandran and Hovy 2002, and also used generalized suffix trees to store potential patterns, Ukkonen, 1995; Geertzen and Van Zaanen, 2004). To optimize the program and to limit the processing time of the algorithm, we extracted only the 50 most frequently occurring patterns that contained the wildcard tokens at least twice. Furthermore, we only considered patterns consisting of at least five words (three words other than the two wildcard tokens), because pilot studies showed that shorter patterns were not specific enough and introduced too much noise in the results.
Step Three: Search the corpus for all occurrences of the selected patterns, where the positions of the wildcard tokens may be taken by any word or set of words. Extract word pairs that fill the wildcard positions.
Step Four: Calculate a ‘score’ for each pattern using knowledge of how often and in what contexts a pattern occurs in the corpus. We define the score S_i as the number of times pattern *i* was found with an already known antonym pair in the wildcard positions, divided by the total number of times it was found in the entire corpus. In other words, the score for a pattern is an estimate of the probability that the pattern co-occurs with an antonym pair. For example, if the pattern *kloof tussen <ant> en <ant>* (‘the rift between <ant> and <ant>’) was found to occur 200 times within the corpus and co-occurred with an antonym pair 50 out of 200 times, its score S_i is 50/200 = 0.25.
Step Five: Discard patterns with a score lower than a set threshold τ. We set the value of τ at 0.02 because pilot testing showed that setting this value lower results in increased noise in the results while setting it higher eliminates too many reasonable patterns.
Step Six: Using the scores of the remaining patterns, calculate the ‘antonymy score’ for each pair that was found in conjunction with these patterns. This score represents the probability that a pair is an actual antonym pair, given how often it occurred with each pattern and the scores of these patterns. Intuitively, it is the probability that it is *not* the case that all evidence for the pair being an antonym pair is false. Formally, it is defined as:

$$A_i = 1 - \prod_j (1 - S_j)^{C_{ij}}$$

where A_i is the score for the i-th pair, S_j is the score for the j-th pattern and C_{ij} is how often the i-th pair was found using the j-th pattern. Consider for example a pair that was found once with pattern P_1 and twice with pattern P_2. If the pattern scores for the patterns are given by $S_1 = 0.6$ and $S_2 = 0.7$, the antonymy score for this pair is calculated as follows[5]:

$$\begin{aligned} A_i &= 1 - (1 - S_1)^1 \times (1 - S_2)^2 \\ &= 1 - (1 - 0.6)^1 \times (1 - 0.7)^2 \\ &= 1 - (0.4)^1 \times (0.3)^2 \\ &= 0.964 \end{aligned}$$

Step Seven: Compile a list of all word pairs found with the patterns and their corresponding antonymy scores. Pairs found with a capital letter or a number (e.g.

Bush-Putin, *50-50*) were filtered from the results because these are unlikely to be antonym pairs. All pairs with a high score were selected as new seeds and the entire process of the identifying patterns and using those to search the corpus were then performed iteratively. We chose to only select pairs with a score equal or above 0.9 as new seeds for the next iteration.

We ran the algorithm for at least six iterations for each of the original seed sets.[6] After the sixth iteration, extracted pairs with an antonymy score higher than 0.6 were compared with the pairs in Dutch WordNet, and with an online dictionary, *Mijnwoordenbok.nl*. Earlier research suggests that available lexical resources are often inconsistent in their identification of antonyms (see e.g. Paradis and Willners 2007 on the inconsistencies of the COBUILD dictionary), and tend to concentrate on the known set of canonical antonyms. To complement evaluation with existing resources we also did a human evaluation. Five native adult speakers of Dutch employed at the Artificial Intelligence department were used as subjects. They were presented with the same randomly ordered list of pairs on a computer screen and asked to classify each pair by choosing the number of the category. Judges were given simple instructions: identify if a pair is a synonym, opposite, correlate, which was explained as sisters in a tree structure, or two elements of the same type, or if the pair fit none of these descriptions (‘none of the above’).

4. Results

4.1. *Extracted pairs*

After the first cycle, 264 pairs were extracted using six seeds and 568 using 18 seeds. The number of found pairs had reached 1189 for six seeds and 1355 for 18 seeds after the final sixth cycle. The overlap between pairs extracted in different sets, 90.4%, or 1075 of the pairs found with six seeds in the last cycle were also found with the set of 18 seeds.

Only pairs with an automatic score ≥0.9 were used as seeds in each consecutive iteration. Since these pairs are likely to be the most reliable results, we will discuss them first.

After the last iteration, 45 extracted pairs in the set of six seeds and 23 extracted pairs in the set of 18 seeds had a score ≥0.9. All, except one, of the pairs that were found in the set of 18 seeds were also found in the set of six seeds. Since corpus was not lemmatized, pairs like *man–vrouw* (‘man’–‘woman’) and *mannen–vrouwen* (‘men’–‘women’) were treated as two different antonym pairs. Among pairs extracted in both sets, there were traditional gradable antonyms like *nieuwe–oude* (‘new’–‘old’) and *grote–klein* (‘large’–‘small’), non-gradable opposites like *links–rechts* (‘left’–‘right’), and opposites like *aanbod–vraag* (‘offer’–‘demand’) and *praktijk–theorie* (‘practice’–‘theory’). Except for the canonical gradable antonyms *arm–rijk* (‘rich’–‘poor’), most of the pairs extracted only in the set of six seeds were opposites that have not been discussed in theoretical literature: *lonen–uitkeringen* (‘salaries’–‘welfares’), *buitenlanders–insiders* (‘foreigners’–‘insiders’), as well as erroneously extracted non-contrastive pairs like *praatjes–praktijk* (‘talks’–‘practice’), *handel–mensenrechten* (‘trade’–‘human rights’), *hem–orkest* (‘him’–‘orchestra’) and even a synonym pair *realiteit–werkelijkheid* (‘reality’–‘actuality’). This shows that there is a trade-off between precision and recall in that using less seeds results in the extraction of a larger number of candidate pairs but consequently more noise whereas using more seeds results in the extraction of a smaller number of more reliable candidate pairs.

Recall that all seeds in both sets were adjectives. Surprisingly, only 15% of the extracted pairs with a score ≥0.9 were also adjectives, while the majority were noun-noun pairs like *man–vrouw* ('man'–'woman') or *moeders–vaders* ('mothers'–'fathers'). Because most of the work in theoretical linguistics has been done on adjectival antonymy, such pairs would be treated as multiple incompatibles rather than 'true antonyms'.

Only two pairs from the original seed sets, namely, *arm–rijk* ('poor'–'rich') and *hoog–laag* ('high'–'low'), were found among pairs with the score ≥0.9 in the sixth cycle. This suggests that our method is successful at extracting new pairs.

4.2. *Evaluation of the extracted pairs*

In the previous section we discussed pairs with a score ≥0.9. In this section we describe general results. In Table 2 we present the number of found word pairs at each score level for each seed set. We also report how many of those pairs co-occurred in sentences in the corpus at a statistically significant level using a χ^2-test, treating $p \leq 0.05$ as significant.

INSERT Table 2

There were more pairs at lower score levels and fewer pairs at higher score levels in both sets of seeds. More pairs were extracted in the set of 18 seeds but their score was lower than 0.6. Manual inspection revealed that these pairs were often poor examples of opposites, so in the rest of the paper we concentrate on pairs with a score of 0.6 or higher. The pairs with a score ≥0.6 co-occurred significantly more often than chance would predict in at least 95.4% of the cases. This supports the idea that significant co-occurrence is one of the prerequisites for antonymy. At higher score levels more pairs were extracted in the set of six seeds. Again, this indicates that the size of the seed set plays a role in that more seeds yield higher precision.

To evaluate the results, we first compared all pairs with a score ≥0.6 with antonyms in the Dutch part of the EuroWordNet (Vossen et al. 1999). Nine out of 197 pairs from the set of six seeds (5%) and ten out of 172 pairs from the set of 18 seeds (6%) are also listed in this lexical resource. One of the reasons why the number is so low could be because the Dutch part of the Euro WordNet does not indicate the relationship of antonymy for adjectives. In addition, many extracted nominal pairs were mutual incompatibles, a category treated in the Euro WordNet as co-hyponyms. But in general, this result illustrates some of the limitations of the presentation of antonyms in Dutch WordNet.

Next, we compared our pairs with a list of antonyms from an online dictionary *Mijn woordenboek*.[7] This list consists of 1184 antonym pairs.[8] Nineteen of the extracted pairs in the set of six seeds (10%) and 20 of the extracted pairs in the set of 18 seeds (12%) were also found in this online resource.

Note that we did not use the widely-known *Van Dale* dictionary for evaluation because the *Van Dale* Dutch-English dictionary (Martin and Tops 1986), the *Van Dale* English-Dutch dictionary (Martin and Tops 1989), as well as a lexical database provided by *Van Dale* were used as a base for the Dutch WordNet.

Finally, all pairs with the score ≥0.6 were also evaluated by human judges. Five participants were presented with the candidate pairs and asked to categorize them as antonyms, synonyms, co-hyponyms or none of those categories. They achieved a

Fleiss's Kappa score of 0.71 for inter-annotator agreement calculated using an online resource (Randolph, J. 2005).[9] A score between 0.61 and 0.8 is considered to indicate substantial agreement, so our results indicate reliable agreement. A pair could be classified as belonging to the same category by all five participants (or unanimously), by three or more participants (that is by the majority) or by two or fewer participants (in which case no majority vote was achieved). Table 3 presents results for the number of pairs classified in each category unanimously by all five participants and by the majority.

INSERT Table 3

The proportion of pairs judged as antonyms was higher in the set of 18 seeds than in the set of six seeds. In both sets there was a difference in the number of pairs categorized as antonyms and co-hyponyms unanimously and by the majority vote. Unanimously, more participants agreed as to the number of antonyms than co-hyponyms. In contrast, by majority, more pairs were judged as co-hyponyms than as antonyms. This suggests that as a group, co-hyponymy is a fuzzier category that includes contrastive but not necessarily mutually exclusive pairs, therefore, it was difficult to decide whether such pairs were co-hyponyms. Antonymy, on the other hand, was defined as a more restrictive category, which lead to a higher agreement among participants. When the two categories are combined, 65.5% of found pairs in the set of six seeds and 68.5% of the found pairs in the set of 18 seeds were judged as contrastive pairs.

As shown in Table 4, human judgments correlated with automatic scoring. The Table presents all pairs from the set of six seeds that were judged as belonging to the same category by three or more participants in relation to automatic scoring.

INSERT Table 4

At each consecutive higher score level, more pairs were classified as antonyms by human judges. The proportion of co-hyponyms, on the other hand, was about the same for every score level. The number of pairs that were judged as antonymous at the score level of ≥0.6 was 28.3%, and it increased to 45.4% at the score level of ≥0.9. The number of pairs that were judged as "none of the above" substantially decreased from 32% at the level of ≥0.6 to 13.6% at the level of ≥0.9. Thus, more pairs with a higher score (above 0.8) were classified as antonyms and fewer were judged as 'none of the above'.

A detailed overview of the kind of pairs we extracted, their translations, the result of the human evaluation, the number of different patterns they occurred in and their syntactic function are all given in Table 6 in the Appendix. For reasons of space, in the Table we present only those pairs that had a score of 0.8 or higher (85 pairs in total).

4.3. *Extracted lexico-syntactic patterns*

All patterns were extracted automatically. At each iteration, the 50 most frequent patterns consisting of at least five elements (including punctuation) were automatically scored. Patterns with a score below the threshold of 0.02 were discarded, patterns with a higher score were used to extract new pairs in the following cycle. A sample of extracted patterns and their types according to Jones (2002) is given in Table 5.

INSERT Table 5

One of the main differences between our automatically extracted patterns and Jones' manually selected patterns (Jones 2002, Jones et al. 2007) is that ours are more specific and diverse, even allowing punctuation marks as part of the pattern, as in '*<ant> en <ant>*,' ('<ant> and <ant>,'). Because we generated many more patterns, the subtle distinctions between specific patterns could be taken into account.

The most frequent types of patterns with high scores were Distinguished, Coordinated and Transitional. But even if specificity of patterns is taken into account, we do not find many of the pattern types identified in Jones (2002). For example, we do not have any variations of the Comparison type (as in '*more <ant> than <ant>*') or the Negated type (such as '*<ant> not <ant>*'). This can be due to the length of such patterns or their frequency. In fact, as has been shown in Muehleisen and Isono (2009), Distinguished, Coordinated and Transitional Antonymy comprise the most frequent types of antonyms identified by patterns in Japanese. Murphy et al. (2009), on the other hand, suggest that in English and Swedish, Comparative type is as frequent as Distinguished and Transitional. This highlights another difference between automatic extraction and evaluation of patterns and their manual identification – the "goodness" of patterns in previous studies did not take into account their frequency as well as the number of times a pattern contained (different) antonym pairs.

Recall that in the first cycle, patterns were identified by means of canonical antonym seed pairs, while in the following cycles, patterns were found using pairs extracted earlier with a scoring ≥0.9. Did the type of pairs used as input affect the kind of patterns we extracted? We can examine this by comparing the patterns found with the seed pairs (cycle one) with patterns found with extracted pairs (cycle six). In both sets, 14 patterns with a score above the threshold were found in cycle one and cycle six. But in general, most of the patterns with a score above the threshold differed at each cycle. A closer look at these patterns revealed that many of them were small variations of the same pattern type. For example, a pattern *kloof tussen <ant> en <ant> <comma>* ('rift between <ant> and <ant> <comma>') was extracted in one cycle while its variation *de kloof tussen <ant> en <ant>* ('the rift between <ant> and <ant>') in another.

Is there a difference between antonyms and co-hyponyms in respect to the types of patterns in which they occur? Pairs with a score ≥0.8 presented in Table 6, all occur in the same type of frequent pattern types (e.g. Distinguished), thus, there seems to be no difference in that respect. On the other hand, we can see that eleven pairs that were classified by human judges as antonymous (40%) and four pairs that were judged as co-hyponyms (12%) occurred in more than five patterns in either set of the seeds. These pairs included adjectival antonyms like *arm–rijk* ('poor'–'rich'), *links–rechts* ('left'–'right'), *blank–zwart* ('white'–'black'), opposites like *man–vrouw* ('man'–'woman'), *aanbod–vraag* ('supply'–'demand') and *werkgevers–werknemers* ('employers'–'employees'), and co-hyponyms like *premier–president* ('prime-minister'–'president'). And most of the pairs that were judged as co-hyponyms occurred in three or fewer patterns. Thus, it seems that the difference between antonyms and co-hyponyms lays not in the type of patterns but rather in the range of patterns they occur in, since co-hyponyms do not occur solely in the contrastive contexts.

5. Discussion

5.1. *Evaluation of the method*

We found that automatically extracted surface level patterns can be used to reliably find antonyms, not just the traditionally identified canonical pairs, and not just antonymous adjectives, even when seeded with adjective pairs.

Table 2 and Table 4 together suggest that the co-occurrence hypothesis, even in combination with frequent appearance in patterns, is not sufficient to distinguish antonyms from non-antonyms, because even among the pairs that co-occcurred statistically significantly more often than chance *and* occurred in reliable patterns there are pairs that are not clear opposites, for example, *bejaarden–vrouwen* ('elderly'–'women'). This result contrasts with the results of Justeson and Katz (1991), but note that they tested sets of antonym pairs in a small set of patterns compared with a set of random adjectives in patterns.

Evaluating the method compared with earlier work on automatic extraction of lexical relations we can conclude that pattern based extraction of antonyms seems to be more difficult than pattern-based hyponymy and meronymy extraction. Taking the majority vote of human judges as the gold standard, we achieve a precision of 28.3 % for pairs with a score ≥ 0.6 and a precision of 45.4% for pairs with a score ≥ 0.9. This is considerably lower than the precision scores presented for automatic hyponym and meronym extraction reported in section 2.3. If we treat co-hyponyms and antonyms together as one class of antonymous pairs we get precision scores of 67.5% and 84.2% respectively, but this is contingent on accepting the class of co-hyponyms as extracted pairs.

With respect to existing lexical resources, the majority of the pairs found are not listed in either Dutch WordNet or the online dictionary *Mijnwoordenboek.nl*. This suggests, as expected by earlier findings for hyponymy and meronymy, that automatic extraction is a useful way to enrich existing resources that are often incomplete and limited due to their reliance on intuition for entries. In our results we see quite clearly the positive contribution the automatic method can make because of its success in extracting even compound words or strings like *informatiearmen–informatierijken* ('information poor'–'information rich') and *werkenden–niet-werkenden* ('working'–'not working'). These pairs are clearly frequently used in opposition in newspaper texts, but are unlikely to be noticed with intuition-based approaches.

The most unexpected result is the large number of noun-noun pairs that the method extracts, contrary to predictions based on the results in Fellbaum (1995). We also extracted several cross-categorical pairs, but they all turned out to be errors, confirm predictions made in Fellbaum (1995) that cross-categorical antonym pairs are unlikely to occur in patterns.. This suggests that knowing whether an extracted pair is cross-categorical or not, can be used to filter errors from the results of pattern-based automatic extraction of antonyms.

5.2. *Noun-noun antonyms vs. co-hyponyms*

The majority of the pairs we found were nouns, and a large portion of these would traditionally be considered co-hyponyms rather than standard antonyms, but they 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 like other antonymous pairs, our results are evidence for treating mutual incompatibles as a subtype of antonyms.

It is crucial to note that not all co-hyponyms have this property, and the 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. Our human judges were often split as to whether a given pair should be classified as co-hyponyms or antonyms. As an example, consider the pair *verdachten–getuigen* ('*suspects'*–'*witnesses'*). Three of the five judges considered the pair to be antonyms, while two considered the words to be co-hyponyms, but considering the contexts, we see that both these functions are possible:

(1) *Journalisten van die media hebben de afgelopen maanden regelmatig nieuws gepubliceerd over het verloop van het onderzoek en de verhoren van verdachten en getuigen.*
In the last months, journalists from the media have regularly published news on the progress of the investigation and on the interrogation of suspects and witnesses.[10]

(2) *Wie lopen het grootste gevaar, de verdachten of de getuigen?*
Who is in the greatest danger, the suspect or the witness?

In (1) the two concepts are used as co-hyponyms; the emphasis is on both types of individuals being subject to interrogation. In (2) the reader is asked to compare suspects to witnesses on the scale of how much danger they are in.

It is unclear if it is the pattern 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. Even pairs that were unanimously judged as co-hyponyms by our judges are frequently used in contexts with a contrastive meaning. Consider the two examples below with the pair *mens–natuur* ('man'–'nature'):

(3) *Niet alleen de natuur, maar ook de mens moest tijdens deze lentefeesten worden gezuiverd om het nieuwe leven toe te laten.*
Not only nature, but also man has to be purified during these spring celebrations in order to let new life in.

(4) *Het gaat immers om een zaak als leefbare, schone steden, en om zoiets als een evenwichtige relatie tussen mens en natuur.*
It is, after all, about having livable, clean cities and about something like a balanced relationship between man and nature.

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, yet we would consider none of these pairs contrastive in the same way that a pair like *binnenstad–suburbia* ('downtown'–'suburbia') is. But our method does not extract such pairs because, unlike antonyms, they do not 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 antonyms. At least for these co-hyponyms, it seems that Fellbaum's (1995) suggestion that antonymy may be an important organizing principle of the mental lexicon for more word classes than just adjectives is at least partially correct.

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 how context-dependent pairs are. Murphy (2006) has suggested that it is the pattern 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 antonyms themselves. For the class of co-hyponyms, identifying the context they are used in seems to be essential to their interpretation.

Also note that the patterns which mark a contrastive or incompatible relationship may be less specific than those for other lexical relations. In the work of Hearst (1992) for hyponym extraction and Berland and Charniak (1999) for meronym extraction it was sufficient to manually identify a set of five (!) lexico-syntactic patterns that consistently identified the targeted lexical relationship with a high precision. For antonym extraction however, even our best patterns do not work this well, and antonym pairs seem to occur in many more contexts, and as a consequence the pairs themselves may also be more context dependent.

5.3. *Limits of pattern-based methods for antonym extraction*

There are a number of limits to the pattern-based approach for antonym extraction. First, the patterns we extract automatically are also qualitatively different from the patterns manually identified in Jones (2002) and used in Jones et al. (2007) to identify canonical antonyms. Our patterns are far more specific, and necessarily so. Recall that Jones et al. (2007) use general patterns where one member of the antonym pair was already filled in. We use patterns to find *both* halves of antonymous pairs. Because of this, patterns that are too general lead to poor precision because they extract many more non-antonyms than candidate antonyms. But this may have an effect on the type of pairs we extract.

Second, although all seeds were adjective pairs, most of the pairs found were expressed by nouns. This result is unexpected given the earlier findings of Justeson and Katz (1991) and Fellbaum (1995). Recall that Justeson and Katz showed that antonymous adjectives are frequently found in parallel aligned syntactic contexts, and Fellbaum argued that this is true only for antonymous adjectives but not nouns or verbs based on searches for sets of cross-categorical antonymous concepts in the Brown Corpus (Kučera and Francis 1967).

It may very well be that verb-verb pairs and cross-categorical antonymous concepts do not occur in recognizable syntactic patterns. Note however that even though we seeded with adjective pairs, noun-noun antonymous pairs frequently occurred in all of the lexico-syntactic patterns that we found. We suspect that this is because the patterns found with adjectives are often in syntactic positions where a noun is also possible, but a verb is not. Our results may differ from Fellbaum (1995) for two reasons. First, we use patterns to search for new pairs, while Fellbaum basically searched for a small set of canonical noun-noun, verb-verb pairs and did not find that they occurred in patterns. Second, we used a corpus that was 72 times as large as the one Fellbaum had available. As part of future work we can see what happens if we use verbal antonym pairs as seeds.

Another limit of the method is that we do not seem to get many more results after a number of cycles. It is true that we did not let the program run until it reached convergence,[11] but we suspect from the way in which the number of useful pairs found increases that quite soon we would fail to obtain new pairs. Researchers have tried to improve the performance of pattern based approaches by using syntactic information to generalize the patterns (e.g. Berland and Charniak, 1999 for meronyms; Snow et al. 2005 for hyponyms, van der Plas and Bouma 2005 for Dutch synonyms) but we wonder if this would help as well for antonyms, because our results all suggest that antonyms are more context-dependent.

The method still extracts a large number of non-opposites, a problem common to most automatic pattern-based methods. We still need to use manual evaluation to sort the useful pairs from the noise. But it is promising to see that our automatic evaluation correlated well with the human evaluation, and in Section 6 on future work we suggest a number of ways in which the automatic evaluation could be improved.

5.4. *Applications of results*

A major problem for clustering techniques that attempt to automatically identify synonyms is that antonyms are also found among the results (Lin et al. 2003). Lin at al. (2003) used two 'patterns of incompatibility', *from X to Y* and *either X or Y*, to separate antonyms from synonyms and other lexically related words. They report a precision of 86.7% and recall of 95%. This suggests that patterns can be used to separate antonyms from synonyms and improve the performance of such NLP tasks as informational retrieval. However, they only used two intuitively selected patterns. In a small pilot study, Lobanova et al. (2009) also tried to use a small list of antonyms to remove antonym errors from automatically acquired synonyms, but only with limited success. Clearly more testing is needed.

A major motivation for doing automatic lexical acquisition in the first place is in order to create or extend lexical resources easily. For a smaller language like Dutch this is particularly necessary. Dutch Wordnet (Vossen et al. 1999) currently does not even include antonymy as a relation between any adjectives, let alone non-canonical ones, and only identifies antonymy as a relation between a small set of mostly canonical nouns and verbs. Our results, even if they require manual verification, could be a useful addition. We found many useful antonyms that are fairly context-independent and would probably be useful if they were incorporated. Pairs such as *schijn–werkelijkheid* ('appearance'–'reality'), *realiteit–utopie*, ('reality'–'utopia') *gevoel–rede,* ('feeling'–'reason') or *daad–word* ('deed'–'word') are good candidates. Of course, given only one member without a context, it is unlikely that the other member would always be the most immediate choice for an opposite; these are not canonical antonyms. But when these pairs occur together, their chance of being in opposition seems to be high. Of course, it is not clear how we should distinguish pairs that we should include in a lexical resource from those that we should not. A quick look at the pairs that scored 0.9 and higher in Table 2 illustrates the problem. It is intuitively clear why each of these pairs has obtained such a high score, because the concepts they denote are frequently used in opposition, but it is not clear if we should add a pair like *salaries–welfare* to a dictionary. Because we use patterns to find pairs, we haven't examined the extent to which the pairs occur outside of the given pattern, and what proportion of these latter uses are actually contrastive. Further research can answer this type of question. Also, as Hearst (1992) noted, for each pair we would wish to add we would have to identify the senses to which the antonym relationship applies, and this can sometimes be more than one. For example, in Dutch WordNet (Vossen et al. 1999) the adjective *groot* ('big') has a sense that means large for a certain type, but also can be used with the sense of 'adult'. For both of these senses *klein* ('small') seems to be an appropriate antonym, but this can hardly be determined automatically in any simple way.

A further application comes from computational discourse research. An extension of the concept of antonymy to non-canonical pairs could be very useful for identifying Contrast relations, and some work along these lines has already been done. Marcu and Echihabi (2002) use word pairs automatically extracted from sentences with the contrastive cue phrase *but* to identify new examples of Contrast relations. Successfully identified relations included pairs like *embargo–legally* on either side of a contrastive marker such as *but*. This pair would be unlikely to be found in any lexical resource as an example of antonymy, especially given that it is a noun-adverb combination. But it does have recognizable contrastive meaning. Spenader and Stulp (2007) used adjective antonym pairs taken from WordNet 2.0 (Fellbaum 1998) and studied how frequently they occurred in *but*-marked and unmarked Contrast relations. They found that surprisingly only 1% of the 218,017 sentences studied contained an adjective antonym pair identified by WordNet. The possible explanation for this low number is that WordNet antonyms are limited to canonical pairs and the indirect antonyms that can be derived from them, while the type of opposition contributing to Contrast relations is likely to include many cases of non-canonical antonyms, and therefore the type of pairs we found would seem to be more useful.

Further, Spenader and Stulp (2007) only looked at antonymous adjectives. However recall that Fellbaum (1995) showed that cross-categorical antonymous concepts also co-occur intrasententially with a frequency greater than chance, but not in parallel syntactic constructions. Fellbaum's (1995) results showed that that many contrastive relationships may be expressed across categories, with perhaps a verb and a noun combination, suggesting that extending the search for antonymous concept pairs (and disregarding syntactic category) as well using non-canonical pairs, might lead to better results. Indeed if Fellbaum (1995) is correct in her claim that antonymous concepts are a major way in which the mental lexicon is organized, then we should expect cross-categorical antonymy to be pervasive in contrastive relationships.

6. Future Work

We would like to do more experiments related to the effect of the number and type of initial seed pairs on the number and quality of the pairs extracted. Right now our only clear result is that it seems with more seed pairs you get a higher precision, but lower recall. But this may have to do with the pairs that were actually in the seed sets, and varying which pairs are in the six-seed set might give different results.

Additional human evaluations of found pairs, varying the possible categories presented, and varying the instructions might also be useful for telling us more about what lexical relation is dominate. It might also be illuminating to ask judges to classify pairs in their sentential contexts as being used contrastively or as co-hyponyms. This might tell us more about which function is more prominent for a given pair, and consequently whether or not the relation should be coded in a lexical resource.

We might also modify the way in which found pairs are evaluated. Currently our method rewards frequent, reliable patterns, and pairs that occurred very frequently even in only one pattern can get a high score. Jones et al. (2007) suggest that the number of different patterns a pair occurs in is indicative of the canonicity of the adjective pair. Even though we are not interested in finding canonical pairs, somehow weighting our results to favor pairs that occur at least in more than one pattern might eliminate a number of incorrect pairs.

There is some additional work that contrasts the contribution of occurring in many patterns to occurring frequently in a very 'good' pattern. Snow et al. (2005) showed that using a classifier trained on a great number of lexico-syntactic patterns was more useful for correctly categorizing hypernym-hyponym pairs than using a smaller set of highly reliable patterns. On the other hand Tjong Kim Sang and Hoffman (2007) showed that when a large quantity of data is used, the most reliable pattern outperforms a combination of patterns for hypernym-hyponym extraction.

It would be interesting to try each of these measures to evaluate which gives more reliable pairs (taking the human evaluation as the gold standard). A first step is to compare the number of different patterns a found pair co-occurs in with the frequency of the pair in all patterns. It would also be interesting to try to identify particularly good patterns and see if a frequent occurrence in these patterns is a more reliable predictor of good opposites compared to frequency alone, especially given that for antonyms we do not find obvious individual patterns that perform at a high level of precision as have been found for hyponymy and meronymy extraction.

Finally, an obvious way to address the problem of distinguishing co-hyponyms from antonyms might be to attempt to do automatic acquisition of co-hyponyms and use the results there to filter our automatically extracted antonyms. In fact, we suspect that noise pairs found in the acquisition of one lexical relation will also be found in the acquisition of other lexical relations, because such pairs happen to be highly correlated words that do not belong to any of the typical lexical relationships. It would be interesting to do a number of experiments combining results from several lexical acquisition attempts to remove noise from the data, extending the ideas of Lin et al. (2003) and Lobanova et al. (2009).

Acknowledgements

Earlier results of this study were presented at the CLIN-2009 meeting (Computational Linguistics in the Netherlands). We would like to thank the audience of the conference for their encouraging questions that helped us to shape our further ideas. We are also grateful to two anonymous reviewers for their comments and suggestions that helped us to improve the paper. We would also like to thank Erik Tjong Kim Sang for useful discussion and comments. Jennifer Spenader's work was supported by the Netherlands Organisation for Scientific Research, NWO, grant no. 016.064.062. All mistakes are our own.

Notes

[1] Note that they studied sets of antonyms. This may have been in part because the one million word Brown Corpus is not large enough to study individual pairs.

[2] Note that these percentages are only approximations, because the set of sentences from which they were extracted were not chosen arbitrarily.

3 Idiomatic Antonymy, where antonyms occurred in idiomatic expressions such as *penny wise and pound foolish,* was a very infrequent class (0.8%) that also did not contain any reliable patterns.

[4] It is not clear if Hearst judged all these pairs to be hypernym-hyponyms. If so, then her precision is 100%.

[5] Note that according to this formula a pair that occurred within a pattern with a maximum score of one at least once, would receive a score of one, becoming a seed pair in the next round. However, since this is an estimated probability, a score of one is too categoric (it suggests that all pairs that ever occur in this pattern are antonyms). To include unseen cases and an error estimation, we lowered the scoring of patterns with a maximum score of one to 0.8. This was an arbitrary choice that is not optimal. Since computational aspects of our work are not the main focus of the paper, a better way of pattern scoring is left for future work.

[6] Another natural stopping point is to let the algorithm run until no more new pairs are found with score over 0.9.

[7] The list of antonyms can be found on the website http://mijnwoordenboek.nl/

[8] We chose to use this online dictionary for evaluation because it had the most comprehensive list of antonym pairs for Dutch that we could find.

[9] http://justus.randolph.name/kappa

[10] All examples in the paper are taken from the CLEF corpus on which the experiments were performed.

[11] This was in part because storing so many patterns was very slow even at six cycles.

References

A. Dictionaries

Martin W.,and Tops, G. A. J. 1986. *Groot woordenboek Engels-Nederlands*. Van Dale Lexicografie. Utrecht.

Martin W.,and Tops, G. A. J. 1989. *Groot woordenboek Nederlands-Engels*. Van Dale Lexicografie. Utrecht.

Roget, P. M. 1952. *Roget's Thesaurus*. London: Sphere Books.

Mijnwoordenboek. 15 July 2009. http://www.mijnwoordenboek.nl/nederlands/.

B. Other literature

Berland M and Charniak, E. 1999. 'Finding Parts in Very Large Corpora.' In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*.

Charles, W.G. and Miller, G.A. 1989. 'Contexts of antonymous adjectives.' *Applied Psycholinguistics* 10.3: 357-75.

Cruse, A. 1986. *Lexical Semantics.* Cambridge: Cambridge University Press.

Deese, J. 1964. 'The Associative Structure of Some Common English Adjectives.' *Journal of Verbal Learning and Verbal Behavior* 3: 347-57.

Deese J. 1965. *The Structure of Associations in Language and Thought*. Baltimore: Johns Hopkins University Press.

Fellbaum, C. 1995. 'Co-occurrence and Antonymy.' *International Journal of Lexicography* 8.4: 281-303.

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

Geertzen, J. and van Zaanen, M. 2004. 'Grammatical Inference using Suffix Trees', In Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI), Athens, Greece.

Hearst, M.A. 1992. 'Automatic Acquisition of Hyponyms from Large Text Corpora.' In: Proceedings of the 14th Conference on Computational linguistics (COLING-92).

Jijkoun, V., G. Mishne and de Rijke, M. 2003. 'Preprocessing Documents to Answer Dutch Questions.' In: Proceedings 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'03), 2003.

Jones, S. 2002. *Antonymy: a corpus-based perspective*, Routledge: London In series: *Routledge Advances in Corpus Linguistics* (edited by Anthony McEnery and Michael Hovey).

Jones, S., M.L. Murphy, C. Paradis and Willners, C. 2007. 'Googling for Opposites – a Web-based Study of antonym Canonicity, *Corpora* 2.2: 129-155.

Justeson, J. and Katz, S. 1991. 'Co-occurrences of Antonymous Adjectives and their Contexts.' *Computational Linguistics* 17: 1-19.

Kučera, H. and Francis, W. N. 1967. *Computational Analysis of present-day American English.* Providence: Brown University Press.

Leech, G. 1974. *Semantics.* New York: Penguin

Lin, D., S. Zhao, L. Qin, and Zhou, M. 2003. 'Identifying Synonyms among Distributionally Similar Words.' In *Proceedings of IJCAI 2003*. Acapulco, Mexico.

Lobanova, A., J. Spenader, T. van de Cruys, T. van der Kleij and Tjong Kim

Sang, E. 2009. '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*, Odense, Denmark.

Lyons, J. 1977. *Semantics*. Cambridge, Cambridge University Press.

Marcu, D. and Echihabi, A. 2002. 'An Unsupervised Approach to Recognizing Discourse Relations.' In *Proceedings of ACL Conference 2002*: 368-375.

Muehleisen, V. and Isono, M. 2009. 'Antonymous adjectives in Japanese discourse.' *Journal of Pragmatics*, 41, 2185-2203.

Murphy, M.L. 2006. 'Antonyms as Lexical Constructions: or, why Paradigmatic Construction isn't an Oxymoron' in D. Schönefeld (ed.) *Constructions All Over: Case Studies and Theoretical Implications*. Special volume of Constructions, SV1–8/2006. Available online at: http://www.constructions-online.de/.

Murphy, M. L., C. Paradis, C. Willners, C. and Jones, S. 2009. 'Discourse functions of antonymy: a cross-linguistic investigation of Swedish and English.' *Journal of Pragmatics* 41: 2159-2184.

Pantel, P. and Pennacchiotti, M. 2006. 'Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations.' In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, 113-120. Sydney, Australia.

Paradis, C. and Willners, C. 2007. 'Antonyms in Dictionary Entries: Methodological Aspects.' *Studia Linguistica* 61.3: 261-277.

Randolph, J. J. 2005. 'Free-marginal Multirater Kappa: An Alternative to Fleiss' Fixed-marginal Multirater Kappa.' Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005.

Ravichandran, D. and Hovy, E. 2002. 'Learning Surface Text Patterns for a Question Answering System.' *In Proceedings of the ACL Conference, 2002*

Snow, R., Jurafsky, D. and Ng, A. Y. 2005. 'Learning Syntactic Patterns for Automatic Hypernym Discovery', *NIPS 17*, Vancouver, Canada.

Spenader, J. & G. Stulp 2007. 'Antonymy in Contrast Relations.' *Seventh International Workshop on Computational Semantics*, Tilburg, 10-12 January, 2007.

Tjong Kim Sang, E. and Hoffman, K. 2007. 'Automatic Extraction of Dutch Hypernym-Hyponym Pairs.' In *Proceedings of CLIN-2006*, Leuven, Belgium, 2007

Ukkonen, E. 1995. 'On-Line Construction of Suffix Trees.' *Algorithmica*, 14, 327-344.

Willners, C. 2001. 'Antonyms in Context. A Corpus-based Semantic Analysis of Swedish Descriptive Adjectives.' Travaux de l'Institut de Linguistique de Lund 40. Lund, Sweden: Lund University Press.

Van der Plas, L. and Bouma, G. 2005. 'Syntactic Contexts for finding Semantically Similar Words.' In Proceedings of *Computational Linguistics in the Netherlands,* (CLIN), 2004.

Vossen, P., Bloksma, L. and Boersma, P. 1999. 'The Dutch Wordnet.' EuroWordNet Paper, July 12, 1999, University of Amsterdam.

Appendix

INSERT Table 6

Tables to be inserted:

Table 1. List of seeds used. The 18 seed set includes the first six seeds plus twelve additional seeds. Pairs marked by the asterisk are also on Jones' list of 56 antonym pairs (2002).

6 canonical seeds	18 canonical seeds	
mooi–lelijk ('pretty'–'ugly')	*droog–nat** ('dry'–'wet')	*blij–verdrietig** ('happy'–'sad')
*arm–rijk** ('poor'–'rich')	*nieuw–oud** ('new'–'old')	*actief–passief** ('active'–'passive')
open–dicht ('open'–'closed')	*hoog–laag** ('high'–'low')	*goed–fout* ('right'–'wrong')
*groot–klein** ('large'–'small')	*koud–warm** ('cold'–'hot')	*dood–levend** ('dead'–'alive')
*snel–langzaam** ('fast'–'slow')	*oud–jong** ('old'–'young')	*zwaar–licht** ('heavy'–'light')
nauw–breed ('narrow'–'broad')	*lang–kort** ('long'–'short')	*hard – zacht** ('hard'–'soft')

Table 2. Number of pairs found at each score level plus percentage of found pairs that also co-occurred statistically significantly more often than chance would predict, listed by initial seed set size.

	6 seeds			18 seeds		
score	***total pairs found***	***signif co-occ***		***total pairs found***	***signif co-occ***	
≥0.2	440	96%	(423)	532	84%	(443)
≥0.4	242	95.4%	(231)	202	97.5%	(197)
≥0.6	197	95.4%	(188)	171	97%	(166)
≥0.8	77	96%	(74)	57	96.4%	(55)
All	1189	90%	(1071)	1355	83.6%	(1133)

Table 3. Majority and unanimous category votes for pairs in both seed sets, five human judges. *Maj* stands for the classification by the majority, *unan* stands for the unanimous classification, and *no maj* stands for no majority.

Type	**6 maj**	**6 unan**	**18 maj**	**18 unan**
antonyms	54 (27.5%)	32 (16.2%)	57 (33.3%)	34 (19.9%)
co-hyponyms	75 (38%)	17 (8.6%)	60 (35%)	15 (8.8%)
synonyms	1 (0.5%)	0	0	0
none	61 (31%)	30 (15.2%)	48 (28%)	26 (15.2%)
no maj	6 (3%)		6 (3.5%)	
Total	197		171	

Table 4. Majority vote for pairs in the six seed set. Percentages are calculated only for pairs where there was a majority vote.

Type	**≥0.6**		**≥0.7**		**≥0.8**		**≥0.9**	
antonyms	28.3%	(54)	27.2%	(49)	35.1%	(26)	45.4%	(20)
co-hyp	39.2%	(75)	39%	(70)	43.2%	(32)	38.7%	(17)
synonyms	0.5%	(1)	0.5%	(1)	1.3%	(1)	2.3%	(1)
none	32%	(61)	33.3%	(60)	20.3%	(15)	13.6%	(6)
Total	191		180		74		44	

Table 5. Some extracted patterns and their type according to Jones (2002).

Functional Type	**Patterns (in Dutch)**	**Patterns (in English)**
Distinguished	kloof tussen <ant> en <ant>	rift between <ant> and <ant>
	de kloof tussen <ant> en <ant>	the rift between <ant> and <ant>
	tussen <ant> en <ant>,	between <ant> and <ant>,
Coordinated	,<ant> en <ant>,	,<ant> and <ant>,
	,<ant> of <ant>,	,<ant> or <ant>,
Transitional	, van <ant> tot <ant>	, from <ant> to <ant>
	van <ant> naar <ant>,	from <ant> till <ant>,
	van <ant> tot <ant>,	from <ant> till <ant>,

Table 6. All pairs with an automatic score ≥0.8 extracted in the 6th cycles of set with six seeds. N.f. = not found, EWN stands for Euro WordNet, Mnw stands for Mijnwoordenboek.nl, N = noun, Adj = adjective, Pro = pronoun, PoS = part of speech, Det = determiner, ant = antonym, syn = synonym, co-hyp = co-hyponym, by maj = by majority, no maj = no majority, N = no, Y = yes, n of p-ns stands for number of patterns.

Dutch	English	score 6 seeds	score 18 seeds	n of p-ns 6 seeds	n of p-ns 18 seeds	by maj	in EWN	in Mwb	PoS
arm - rijk	poor – rich	1	1	11	14	ant	N	Y	Adj – Adj
mannen - vrouwen	men – women	1	1	11	14	ant	Y	N	N – N
lonen - uitkeringen	salaries – welfares	1	0.86	6	4	co-hyp	N	N	N – N
man - vrouw	man – woman	1	1	17	19	ant	Y	N	N – N
gekozenen - kiezers	elected – electorate	0.99	0.99	3	2	ant	N	N	N – N
burgers - politiek	citizens – policy	0.99	0.99	7	7	co-hyp	N	N	N – N
burger - politiek	citizen – policy	0.99	0.99	7	6	co-hyp	N	N	N – N
arme - rijke	poor – rich	0.99	0.99	4	5	ant	N	Y	Adj – Adj
blank - zwart	white – black	0.99	0.99	7	6	ant	N	N	Adj – Adj
burger - overheid	citizen – government	0.99	0.99	4	2	co-hyp	N	N	N – N
links - rechts	left – right	0.99	0.99	11	13	ant	N	Y	Adj – Adj
loonontwikkeling - uitkeringen	salary growth – welfares	0.99	n.f.	3	n.f.	none	N	N	N – N
dollar - peso	dollar – peso	0.99	n.f.	3	n.f.	co-hyp	N	N	N – N
gekozene - kiezer	elected – electorate	0.98	0.98	3	3	ant	N	N	N – N
mens - natuur	people – nature	0.98	n.f.	4	n.f.	co-hyp	N	N	N – N
bestuur - burger	governing body – citizen	0.98	0.97	2	3	co-hyp	N	N	N – N
moeders - vaders	mothers – fathers	0.97	0.97	3	3	co-hyp	N	N	N – N
hoog - laag	high – low	n.f.	0.97	n.f.	3	ant	N	Y	Adj – Adj
aanbod - vraag	offer – demand	0.97	0.96	6	5	ant	N	Y	N – N
praktijk - theorie[1]	practice – theory	0.97	0.97	4	7	ant	N	Y	N – N
kiezer - overheid	elector – government	0.97	n.f.	2	n.f.	co-hyp	N	N	N – N
nieuwe - oude	new – old	0.96	0.96	8	9	ant	N	Y	Adj – Adj
maatschappij - wetenschap	society – science	0.96	n.f.	2	n.f.	co-hyp	N	N	N – N
perceptie - werkelijkheid	perception – reality	0.96	n.f.	1	n.f.	co-hyp	N	N	N – N
christen - jood	Christian – Jew	0.96	n.f.	1	n.f.	co-hyp	N	N	N – N
hem - orkest	him – orchestra	0.96	n.f.	1	n.f.	none	N	N	Pro-N

[1] The Latin word form "practicus" is listed with "theoreticus" as a near-antonym and since 'theorie" and "praktijk" are non-Latin versions of the same pair we have counted this as a variant. The word "praktijk" with the intended meaning of "practice" is not an entry.

binnenstad - suburbia	city center – suburbia	0.96	n.f.	1	n.f.	ant	N	N	N– N
praatjes - praktijk	talks – practice	0.96	n.f.	1	n.f.	none	N	N	N– N
realiteit - utopie	reality – utopia	0.96	n.f.	1	n.f.	ant	N	N	N– N
buitenstaanders - insiders	foreigners – insiders	0.96	n.f.	1	n.f.	ant	N	Y	N– N
bedoeling - effect	intent – result	0.96	n.f.	1	n.f.	no maj	N	N	N– N
realiteit - werkelijkheid	reality – actuality	0.96	n.f.	1	n.f.	syn	N	N	N – N
doden – levenden[2]	dead – alive (pl)	0.95	0.94	3	3	ant	Y	Y	N – N
armen – rijken	poor – rich (pl)	0.95	0.93	2	2	ant	N	Y	N – N
kiezer - politiek	elector – policy	0.95	0.93	2	2	co-hyp	N	N	N – N
handel - mensenrechten	trade / human rights	0.95	n.f.	2	n.f.	none	N	N	N – N
bevolking - politiek	population – policy	0.94	0.93	2	2	co-hyp	N	N	N – N
aankoopgegevens - adres	purchase information – address	0.94	n.f.	1	n.f.	none	N	N	N – N
atoomkernen - elektronen	atom cores – electrons	0.94	n.f.	1	n.f.	co-hyp	N	N	N – N
bestuurders - burgers	politicians – citizens	0.94	0.92	1	1	co-hyp	N	N	N – N
gemiddelde - uitkeringen	average – welfares	0.94	n.f.	3	n.f.	none	N	N	N – N
bruto - netto	gross – net	0.93	0.91	3	3	ant	N	N	N – N
niet-werkenden - werkenden	not-working – working (pl)	0.91	0.83	6	5	ant	N	N	N – N
grote - kleine	large – small	0.91	0.88	7	6	ant	N	Y	Adj –Adj
cao-lonen - uitkeringen	‘cao’-salaries – welfares	0.90	n.f.	2	n.f.	co-hyp	N	N	N – N
have-nots - haves	have-nots – haves	0.90	0.89	5	6	ant	N	N	N – N
oost - west	east – west	0.89	0.89	4	5	ant	N	N	N – N
militaire - politieke	military – political	0.88	0.85	6	5	co-hyp	N	N	Adj – Adj
ambtenaren - werknemers	civil servant – employee	0.87	0.86	1	1	co-hyp	N	N	N – N
banenpoolers - huidige	job-poolers – current	0.87	0.86	1	1	none	N	N	N-Adj
gesproken - gezongen	spoken – sung	0.87	0.86	1	1	co-hyp	N	N	Adj – Adj
weduwen - weduwnaars	widow – widower	0.87	0.86	1	1	co-hyp	Y	N	N – N
gepensioneerden - slapers	retired – overnight guests	0.87	0.86	1	1	none	N	N	N – N
deeltijdwerkers - werknemers	part-time workers – employees	0.87	0.86	1	1	co-hyp	N	N	N – N
scheidsrechters - spelers	referees – players	0.87	0.86	1	1	co-hyp	N	N	N – N
gehuwden - ongehuwd	married couples – unmarried	0.87	0.86	1	1	ant	N	N	N – N
afnemers - zo	customer – therefore	0.87	0.86	1	1	none	N	N	N – Adv
buitenlandse - nationale	foreign – national	0.87	0.86	1	1	ant	N	N	Adj – Adj
gehuwden - samenwonenden	married people – couples living together	0.87	0.86	1	1	co-hyp	N	N	N – N

[2] Included is the noun “dood” (death) and “leven” (life”). Dutch Euro WordNet only marks antonymy for noun and verb relations.

eigen-woningbezitters- huurders	homeowners – tenants	0.87	0.86	1	1	co-hyp	N	N	N – N
openbare - vrije	public – free	0.87	0.86	1	1	no maj	N	N	Adj- Adj
loon - uitkering	salary – welfare	0.86	0.81	2	3	co-hyp	N	N	N – N
werkgevers - werknemers	employer – employee	0.85	0.8	5	5	ant	N	N	N – N
commerci - publieke	commercial companies – public	0.84	0.82	4	4	no maj	N	N	N – Adj
minimumloon - sociale	minimum wage – social	0.83	n.f.	2	n.f.	none	N	N	N – Adj
bestuurder - burger	manager – citizen	0.83	n.f.	2	n.f.	co-hyp	N	N	N – N
premier – president	prime minister – president	0.83	0.83	5	6	co-hyp	N	N	N – N
moeder-vader	mother – father	n.f.	0.83	n.f.	4	ant	N		N – N
maatschappij - politiek	society – politics	0.83	n.f.	2	n.f.	co-hyp	N	N	N – N
stijging - uitkeringen	increase – welfare	0.83	n.f.	1	n.f.	none	N	N	N – N
ambtenarensalarissen - loonontwikkeling	government employee salaries – salary development	0.83	n.f.	1	n.f.	none	N	N	N – N
loonkosten - netto	salary costs – netto	0.82	n.f.	2	n.f.	no maj	N	N	N – N
christenen - moslims	christens – muslims	0.82	n.f.	5	n.f.	co-hyp	N	N	N – N
gewonden - vrouwen	wounded – women	0.82	0.8	1	1	none	N	N	N – N
eken - keybord-orkest	oaks – keyboard ochestra	0.82	0.8	1	1	none	N	N	N – N
familie - ouders	family –parents	0.82	0.8	1	1	co-hyp	N	N	N – N
getuigen - verdachten	witnesses – suspects	0.81	0.8	1	1	ant	N	N	N – N
dat - werkelijkheid	that – reality	0.8	n.f.	2	n.f.	none	N	N	Det – N
bejaarden - vrouwen	elderly – women	n.f.	0.8	n.f.	1	none	N	N	N – N
joden - volwassenen	jews – adults	n.f.	0.8	n.f.	1	none	N	N	N – N
echtgenoot - ouders	spouse – parents	n.f.	0.8	n.f.	1	co-hyp	N	N	N – N
bejaarden - moeders	elderly – mothers	n.f.	0.8	n.f.	1	none	N	N	N – N
gezinnen - vrouwen	family – women	n.f.	0.8	n.f.	1	co-hyp	N	N	N – N
kerk - keuken	church – kitchen	n.f.	0.8	n.f.	1	none	N	N	N – N
droom - werkelijkheid	dream-reality	0.8	n.f.	2	n.f.	ant	N	N	N – N