What implementation and translation teach us: the case of semantic similarity measures in wordnets

Marten Postma
Utrecht University
Utrecht, Netherlands.
martenp@gmail.com

Piek Vossen
VU University Amsterdam
Amsterdam, Netherlands
piek.vossen@vu.nl

Abstract

Wordnet::Similarity is an important instrument used for many applications. It has been available for a while as a toolkit for English and it has been frequently tested on English gold standards. In this paper, we describe how we constructed a Dutch gold standard that matches the English gold standard as closely as possible. We also re-implemented the WordNet::Similarity package to be able to deal with any wordnet that is specified in Wordnet-LMF format independent of the language. This opens up the possibility to compare the similarity measures across wordnets and across languages. It also provides a new way of comparing wordnet structures across languages through one of its core aspects: the synonymy and hyponymy structure. In this paper, we report on the comparison between Dutch and English wordnets and gold standards. This comparison shows that the gold standards, and therefore the intuitions of English and Dutch native speakers, appear to be highly compatible. We also show that our package generates similar results for English as reported earlier and good results for Dutch. To the contrary of what we expected, some measures even perform better in Dutch than English.

1 Introduction

Various methods have been proposed in the past for measuring similarity between words using Princeton WordNet (Fellbaum, 1998). Some of these methods (*path* (Rada et al., 1989), *lch* (Leacock and Chodorow, 1998), *wup* (Wu and Palmer, 1994), *res* (Resnik, 1995), *lin* (Lin, 1998), *jcn* (Jiang and Conrath, 1997), among others) were implemented in the WordNet::Similarity package (Pedersen et al., 2004). WordNet::Similarity [1] has become an important instrument for measuring similarity between any set of words in a language but also for testing the performance of wordnet as a database of synonymy and semantic relations. The toolkit was used to evaluate the different measures against a gold standard of English words created by Rubenstein and Goodenough (1965) and Miller and Charles (1991). The evaluation results tell us something about the capacity of WordNet to mimic human judgements of similarity but also about the different methods in relation to each other.

Unfortunately, WordNet::Similarity only works for the Princeton WordNet released in its proprietary format and not wordnets in other languages in other formats, such as Wordnet-LMF (Vossen, Soria and Monachini, 2013). Furthermore, no gold standard exists for Dutch, the language that we study. In this paper, we describe a re-implementation of the WordNet::Similarity toolkit that can read any wordnet in Wordnet-LMF format to apply the 6 wordnet similarity algorithms. This toolkit makes it possible to carry out similarity measures across different wordnets within the same language and across different languages. This is especially useful if the wordnets were created independently using their own semantic hierarchy. We also created a gold standard in Dutch that is comparable with the gold standard in English. We tried to recreate the process through which the English gold standard was created as much as possible. Since it was not clear what instructions were given exactly to the human scorers, we decided to create a number of additional gold standards that are more explicit about the difference between relatedness, similarity and the assumed meaning of the words to be com-

[1] see http://wn-similarity.sourceforge.net/

pared. In total 6 different gold standards have been created. Using these gold standards, we first show that the 6 Dutch gold standards are very similar and that the English and Dutch gold standards are highly compatible. Secondly, we demonstrate that the performance of the Dutch wordnet is higher than the reported performance for English. There are also some differences in the results which can be explained to some well-known differences in the hierarchical organization of the Dutch and English wordnets.

The paper is structured as follows. In the next section, we describe related work. Section 3 explains how we created the Dutch gold standard and section 4 the WordnetTools implementation of the similarity functions. In section 5, we report the results using the Dutch wordnet Cornetto 2.1 (Vossen et al., 2013).

2 Related work

The notion of similarity is central to WordNet through the relations synonymy and hyponymy. Synsets group words that can be exchanged in contexts and thus have more or less the same denotational domain. Hyponymy groups these synsets according to a shared semantic aspect and thus defines another type of similarity. Words that do not share a synonymy relation and synsets that do not share a hyponymy relation are not necessarily disjoint but the things they can refer to are less likely to be considered similar. Words and synsets that have other relations than synonymy and hyponymy respectively, e.g. part-whole or causal relations, are most likely not similar but strongly related. This difference is dubbed the 'tennis-phenomenon' in Fellbaum (1998) : where *tennis ball*, *player*, *racket* and *game* are closely related but all very different things. Since WordNet dominantly consists of synonymy and hyponymy relations, it more naturally reflects similarity than relatedness.

Since the first release of WordNet, researchers have tried to use it to simulate similarity. Except for the *lesk* (Lesk, 1986), *vector* (Patwardhan and Pedersen, 2006), and *vector pairs* (Patwardhan and Pedersen, 2006) algorithms, these measures are all based on synonymy and hyponymy.

Another approach to measure similarity across different languages is described by Joubarne and Inkpen (2011). The aim of their paper is to show that it might be possible to use the scores from the English gold standards in other languages, hence making it unnecessary to create gold standards with human-assigned judgements in every single language. In order to show this, they used an existing gold standard for German, which is a translation of the gold standard by Rubenstein & Goodenough with human-assigned scores. For French, they used an existing French translation of the English gold standard by Rubenstein & Goodenough, and asked French native speakers to rate the similarity of meaning for each word pair in the dataset. Moreover, they used two measures of similarity to also rate the similarity of meaning of the translation of the original dataset, which are Point-wise mutual information and second order co-occurence Point-wise mutual information for which the Google n-gram corpus was used. They then compared the output from the similarity measures to the language specific gold standards and to the original scores collected by Rubenstein & Goodenough. The difference between these correlations was relatively small, which is why they claim that it is possible to use the original scores from the English gold standard in other languages.

Besides Joubarne and Inkpen (2011), other studies have made an effort to translate the original datasets by Rubenstein & Goodenough and by Miller & Charles. Hassan and Mihalcea (2009) translated these datasets into Spanish, Arabic, and Romanian. For Spanish, native speakers, who were highly proficient in English, were asked to translate the datasets. They were asked not to use multi-word expressions. They were asked to take into account the relatedness within a word pair for disambiguation. In addition, they were allowed to use so-called replacement words to overcome slang or if words were culturally dependent. They then asked 5 participants to rate the Spanish word pairs. A sixth person evaluated the translation. Because of the fact that the Pearson correlation with the original datasets was 0.86, only one translator translated the datasets into Arabic and Romanian. Finally, Gurevych (2005) translated the datasets into German. However, no instructions, as to how it was done, were provided.

3 Dutch gold standard

We would like to see whether the similarity intuitions of Dutch speakers are the same as the English speakers. We also want to known if the Dutch wordnet Cornetto, which was built independently of the English WordNet, would perform in the same way as the English WordNet using the same similarity measures and against a comparable gold standard. For that, we need to create a Dutch gold standard. We opted to translate the gold standards by Rubenstein & Goodenough (65 word pairs) and by Miller & Charles (30 word pairs). Because the words used by Miller & Charles are a subset of the words used by Rubenstein & Goodenough, and because words are used more than once in both experiments, there are only 49 unique words used in both experiments. In addition, Miller & Charles made one change to the dataset by Rubenstein & Goodenough. Whenever Rubenstein & Goodenough used the word *cord*, Miller & Charles uses the word *chord*.

Inspired by Hassan and Mihalcea (2009), the following general procedure is followed in the translation of the 49 words: [2]

1. The first step is to disambiguate the English word forms. The English experiments present a word form and not a specific concept the word refers to. The results from human judgement provide a good indication as to which concept in WordNet is meant.

2. Following the results in 1, a Dutch translation is chosen for each word.

3. In addition, it is checked whether the relative frequency of the Dutch and English words are in the same class of relative frequency. This is done in order to make sure that there are no outliers. A translation is an outlier when its relative frequency deviates significantly from the original word.

We will now discuss each step of the general procedure in more detail. The first step consists of disambiguating the 49 English words. For example, WordNet lists two senses for the word *asylum*:

1. 'a shelter from danger or hardship'

2. 'a hospital for mentally incompetent or unbalanced person'

In the results of Miller & Charles and Rubenstein & Goodenough, we observe that the correlation with *madhouse* is very high. Hence, the second sense as listed in WordNet is chosen for *asylum*. The same procedure is applied to all other words.

The next step is to translate all English words into Dutch. One of the difficulties we encountered was the case in which two synonyms were used in English, but no two contemporary Dutch synonyms were available. When we encountered such a problem, we opted to replace the English synonyms with two Dutch synonyms that were closely related to the English synonyms. For example, due to the fact that there is only one common Dutch word *haan* "male chicken" for the English synonyms *cock* and *rooster*, we opted to replace these two words by *kip* "female chicken" and *hen* "female chicken", the two Dutch words for female chickens.

In addition, the relative frequencies of the English word and its translation were checked. In order to calculate relative frequencies of the English words, the English sense-tagged corpus SemCor (Miller et al., 1993) was used. For Dutch, such a resource was not available. We are aware of the fact that the Dutch sense-tagged corpus DutchSemCor (Vossen et al., 2012) exists. However, an effort was made to provide an equal number of examples for each meaning in this corpus. Although this is very useful for WSD-experiments, this makes this corpus less useful for Information Content calculations. Therefore the frequencies of the lemmas in the Dutch corpus called SoNaR (Oostdijk et al., 2008) were used. It was checked whether or not the English word and its Dutch counterpart were located in the same class of relative frequency. A word is placed in the category **high** if its relative frequency is higher than 0.05%, **middle** if its relative frequency is between 0.015% and 0.05% and **low** if its relative frequency is lower than 0.015%. If two words are located in the same relative frequency class, the pair receives the value True, else False. If no frequency data was available for a word, the value of the pair was set to True. Eight word pairs received the value False. Since this step was performed to remove outliers, we claim this to be acceptable.

The Dutch translation was then used to reproduce the English experiments by Miller & Charles and Rubenstein & Goodenough. Since the instructions concerning *Similarity of meaning* are un-

[2] We made an effort to compare the polysemy of the English word and its translation. However, English words in WordNet tend to have many more meanings than words in Cornetto. In addition, Dutch words often only refer to one specific part-of-speech, whereas English words often have noun and verb meanings. Because of these differences, we decided not to use this means of comparison in our translation procedure.

clear in the original experiments, we reproduced each experiment with three different kinds of instructions, which are *stressing similarity aspects*, *stressing relatedness aspects*, and *no instructions*. These instructions were explained to the participants by an example of each value that could be assigned to a word pair and a general description. The WordSimilarity-353 Test Collection (Finkelstein et al., 2002) was used to obtain example word pairs for each value that could be assigned to a word pair. This dataset contains two sets of English word pairs with similarity scores assigned by humans. The first set of this collection contains 153 word pairs, with their scores, from 0 to 10, assigned by 13 subjects. In addition, participants were asked to rate the word pairs on similarity. From this set, examples were chosen *stressing similarity aspects*. The second set contains 200 word pairs, with human-assigned scores, from 0 to 10, by 16 subjects. In this case, participants were asked to rate the word pairs based on relatedness. From this set, examples were chosen *stressing relatedness aspects*. Each word pair that was chosen to serve as an example word pair was translated into Dutch. For *stressing similarity*, participants were asked to indicate to what degree two words could replace each other. For example, if two words were interchangeable, they were told to assign the highest value. They were instructed to assign a lower value to a word pair like *aardappelmesje* 'potato peeler' & *mes* 'knife', because *mes* 'knife' can be used instead of *aardappelmesje* 'potato peeler', but not the other way around. For stressing *relatedness aspects*, participants were asked to focus on how likely it is that words occur in the same situation. For example, it is very likely that *computer* 'computer' & *internet* 'internet' occur in the same situation together, whereas this is less likely the case for *komkommer* 'cucumber' & *professor* 'professor'. Finally for the *no instructions* case, the interpretation was left to the participant, except that we indicated that synonyms resulted in the highest score. Combining the two English experiments with the three different kinds of instructions thus yielded six different sets. For convenience, we will use abbreviations to refer to the six experiments. The abbreviation *Mc* will be used for the translation of the dataset by Miller & Charles. *Rg* will be used for the translation of the dataset by Rubenstein & Goodenough. In addition, the three kinds of instructions will be abbreviated in the following way: *No* for no instruction, *Sim* for similarity, and *Rel* for relatedness. By combining the abbreviations, we can refer to each of the six experiments. For example, *McSim* means that the translation of the dataset by Miller & Charles is meant with the instruction *similarity*. Pupils from five Dutch high schools participated. The pupils's age ranged from 16 to 18 years. Their level of education was one the two highest levels of Dutch secondary education, called *HAVO* and *VWO*. Numbers of participants per experiment were: 40 for *McNo*, 40 for *McRel*, 52 for *McSim*, 26 for *RgNo*, 42 for *RgSim*, and 40 for *RgRel*. The difference between the results of the different instructions turned out to be neither significant, nor systematic. We thus assume that the instructions have not been effective to override the basic intuition of the participants.

4 WordnetTools

WordnetTools is a reimplementation of the WordNet::Similarity package in Java1.6 that can read any wordnet in WordNet-LMF format to apply the major similarity functions: Path, Jiang & Conrath, Leacock & Chodorow, Lin, Resnik, Wu & Palmer (see above). The similarity functions can be tuned using various parameters:

–lmf-file Path to the wordnet file in LMF format. A few other formats are also supported.

–pos (optional) part-of-speech filter, values: n, v, a.

–relations (optional) file with relations used for the hierarchy, if not selected a standard set of relations is used: hypernym, has_hypernym, has_hyperonym, near_synonym, eng_derivative, xpos_near_synonym, xpos_near_hyperonym, xpos_near_hypernym.

–input File with pairs to be compared on single lines, separated with backward slash.

–pairs The type of input values: "words" or "synsets" or "word-synsets pairs"

–method leacock-chodorow, resnik, path, wu-palmer, jiang-conrath, lin or all.

–depth Optional: a fixed value for average depth can be given.

–subsumers Path to a file with subsumer frequencies, required for resnik, lin, jiang-conrath or all.

–separator Token for separating input and output fields, default is TAB.

The above options can be used to configure the experiments and the way similarity is calculated. The graph through which words and synsets are compared can be restricted by selecting the part-of-speech or specifying a certain set of relations. The internal data structure treats the result as a graph without further distinguishing the type of relations. It is for example possible to accept strict hypernym relations and looser relations such as near_synonym, xpos_hyperonym and xpos_near_synonym relations for all parts of speech. The toolkit will then build a graph in which synsets are connected through any of these relations.[3] Against such a graph, words such as *transport* as a verb and *transportation* and *transport* as nouns will get scores similar to co-hyponyms. The more relations are included, such as role and causal relations, the more the graph will measure relatedness instead of similarity. For the purpose of this paper, we configured the settings so that graph is most similar to the hierarchical structure of the English WordNet. We thus only used the has_hypernym and has_hyperonym relations.

The toolkit can handle tangled structure as a result of e.g. multiple hypernyms. In case of multiple hypernyms, all possible paths are calculated and given back as the set of paths through the graph. Similarly, if a word has multiple senses, we generate all possible paths for each sense. When comparing two words, we compare all paths of one word with all paths of another word and calculate the similarity score to the specified metrics using each pair of paths. In the end, we keep the paths with the best result. Note that for measures that use information content this is not always the shortest path.

In addition to the similarity API, the toolkit also provides a number of auxiliary functions, for example to determine the average or maximum depth for a wordnet per part-of-speech. WordnetTools is freely available under the GPLv3 license and can be downloaded from: http://wordpress.let.vupr.nl/software/wordnettools/. The package includes the Dutch and English gold standards, as well as the English WordNet in Wordnet-LMF format and the English SemCor frequencies in the proper import format. It also includes the results of the Dutch and English evaluation. The Cornetto wordnet is not included since it is restricted by license. A free research license can be obtained from the Dutch centre for language technology (TST-centrale[4]). However, we will release an open-source version of the Dutch wordnet, which will be included in the package when released. Also the SoNaR word frequencies can be obtained from the TST-centrale. The SoNaR word frequencies have been converted to the hypernym frequencies as described by Resnik, by averaging frequencies over the senses of a word and transferring these to the hypernyms (and further up the hierarchy). These derived hypernym frequencies are also included in the package.

5 Results

Three evaluations have been run to compare the similarity measures across wordnets and across languages. We start by comparing the Dutch to the English gold standards, followed by an evaluation of the comparison between the Dutch gold standards and the similarity measures. Finally, we try to replicate the English experiment by Pedersen (2010) using English Wordnet-LMF and Wordnet-Tools. [5]

5.1 The Dutch gold standard with the English gold standard

The first evaluation that we carried out is the comparison between the English gold standards and their Dutch translations. Since we have an equivalence relation between most of the words, we can compare the rankings of the Dutch and English native speakers. In the evaluation, we left out the word pairs in which a word had not been directly translated, which was the case for word pairs like *cock* and *rooster*. Table 1 presents the evaluation:

Dutch Gold standard	**Spearman ρ original dataset**
McNo	0.88
McSim	0.86
McRel	0.89
RgNo	0.93
RgSim	0.93
RgRel	0.93

Table 1: Evaluation of the comparison between the English gold standards and their Dutch translations.

[3] If bi-directional relations are used in the wordnet, only one of these should be chosen. If not, the path-construction can be terminated by direct circularity of the bi-directional relations.

[4] see `http://tst-centrale.org/`

[5] A github has been created to make it possible to replicate the results in this section. The url to this github is `https://github.com/MartenPostma/PostmaVossenGWC2014`

The results show that the English and Dutch intuitions concerning *Similarity of meaning* are very similar. The range of the Spearman ρ correlation is between 0.86 and 0.93. It also shows that there is little difference across the different Dutch gold standards. The gold standard with similarity instructions (Sim) performs a bit lower on the Miller & Charles set but this difference disappears on the Rubenstein & Goodenough set.

5.2 Comparing Cornetto with the Dutch gold standard

The second evaluation consists of comparing the Dutch gold standards to the output of the similarity measures as calculated in Cornetto using the WordNetTools. We used the following settings to run WordNetTools:[6]

–lmf-file Path to Cornetto in LMF format

–pos no pos-filter was used

–relations has_hypernym, has_hyperonym,

–input path to Dutch gold standards

–pairs "words"

–method all.

–depth 15

–subsumers path to subsumers from the SoNaR word-frequencies

Table 2 presents the results for the different measures on the Dutch gold standard.

SM	McNo	McRel	McSim	RgNo	RgRel	RgSim
path	0.840	0.796	0.856	**0.783**	**0.720**	**0.777**
lch	0.840	0.796	0.856	**0.783**	**0.720**	**0.777**
wup	0.806	0.766	0.831	0.770	0.704	0.769
res	0.765	0.737	0.785	0.720	0.669	0.719
jcn	**0.852**	**0.797**	**0.891**	0.525	0.488	0.512
lin	0.838	0.779	0.880	0.531	0.495	0.520

Table 2: The Spearman ρ is shown by comparing all six similarity measures to all six gold standards.

In general, the results show that all six semantic similarity measures correlate well with the gold standards. *Jcn* correlates best with the translation of the Miller & Charles' gold standards, whereas this is true for *path* and *lch* for the Rubenstein & Goodenough' gold standards. Finally, there is a significant difference between the performance of the measures *lin* and *jcn* when compared to the Miller & Charles' gold standards or the Rubenstein & Goodenough' gold standards. The gold standards are however too small to derive any conclusions from these differences. Larger more representative experiments are needed for that.

5.3 Replication English with Wordnet-LMF and WordnetToolkit

The final evaluation consists of comparing the WordNet::Similarity package to the WordnetTools. This is mainly done to verify if the implementations of the semantic similarity measures are compatible across the packages, i.e. can we reproduce the results of WordNet::Similarity with the original WordNet database with WordnetTools with the WordnetLMF version of the English WordNet. In order to do this, we compare the correlations that Pedersen (2010) reports when calculating the correlations between the original gold standards and the scores from the six similarity measures using WordNet::Similarity to the same procedure but using the WordNetTools to compute the similarity scores.

We used the following settings for WordNetTools:[7]

–lmf-file Path to WordNet in LMF format

–pos no pos-filter was used

–relations has_hypernym, has_hyperonym,

–input path to English gold standards

–pairs "words"

–method all.

–depth 19

–subsumers path to subsumers using SemCor

Table 3 presents the results. The second and third column present the correlation as reported by Pedersen and by our package, respectively, for the gold standard by Miller & Charles, followed by the difference between the two correlations. The other columns presents the same scores for the gold standard by Rubenstein & Goodenough.

SM	McPed	McWT	diff	RgPed	RgWT	diff
path	0.68	0.72	-0.04	0.69	0.78	-0.09
lch	0.71	0.72	-0.01	0.70	0.78	-0.08
wup	0.74	0.74	0.00	0.69	0.78	-0.09
res	0.74	0.75	-0.01	0.69	0.76	-0.07
jcn	0.72	0.65	0.07	0.51	0.56	-0.05
lin	0.73	0.67	0.06	0.58	0.60	-0.02

Table 3: Comparison of the results by Pedersen (2010) and the replication of these results using Wordnet-LMF and the WordnetToolkit

[6] The depth parameter is set to 15, which is mainly relevant for the measure *lch*, which requires the maximum depth of the taxonomy in which the synsets are located. In the case for nouns in Cornetto, this value is 15. For more information, we refer to section 6.

[7] The depth parameter is set to 19, For more information, we refer to section 6.

The results show that for both gold standards, we approach the correlations that are reported by Pedersen (2010), but that there are probably still differences in the implementation of the measures that lead to different output values.

6 Discussion

Three main points stand out in the results. Firstly, the correlations between the English and Dutch gold standards are very high. Given the fact that this was also the case for the Spanish and English intuitions, as discussed by Hassan and Mihalcea (2009), it might be the case the people with different mother tongues have a shared sense of *similarity of meaning*. It should be noted that all speakers from the different languages share a similar Western background. Secondly, the results for Dutch are generally higher than for English. We have no clear explanation for this difference. We know that the Dutch hypernym structure for nouns is more shallow than the English hierarchy. Evidence for this claim can be found in table 4, which shows the noun synset depth distribution for both Cornetto and Princeton WordNet:

	Cornetto		**Princeton WordNet**	
D	NoS	P	NoS	P
0	833	1,26%	1	0,00%
1	8	0,01%	59	0,06%
2	2138	3,23%	3286	3,45%
3	2748	4,16%	3943	4,14%
4	7476	11,31%	3222	3,38%
5	15896	24,04%	3186	3,34%
6	15304	23,15%	5951	6,24%
7	8902	13,46%	10474	10,99%
8	4441	6,72%	18071	18,96%
9	2603	3,94%	16049	16,84%
10	2211	3,34%	12313	12,92%
11	1858	2,81%	7984	8,38%
12	1228	1,86%	4714	4,95%
13	406	0,61%	2634	2,76%
14	66	0,10%	1511	1,59%
15	3	0,00%	917	0,96%
16	0	0,00%	468	0,49%
17	0	0,00%	345	0,36%
18	0	0,00%	165	0,17%
19	0	0,00%	30	0,03%
Total	66121	100%	95323	100%

Table 4: Synset frequency and percentage of total number of synsets is shown for every depth value in Cornetto as well as WordNet. **D** abbreviates ‘depth’, **NoS** ‘number of synsets’ and **P** 'percentage of total number of synsets’.

Table 4 shows that the most frequent depth in Cornetto is 5, whereas this is 8 for Princeton WordNet. In addition, if we calculate the average noun depth in both lexical semantic databases based on the numbers in table 4, we observe that the average noun synset depth in Cornetto is 6.03 and 8.38 for Princeton WordNet. A flatter hiearchy may lead to a more rough but more uniform measure across different parts of the hiearchy. Neverthless, it does not explain the higher correlation with human intuitions. We also know that the Dutch wordnet has more multiple hypernyms. Table 5 provides evidence for this claim:

	Cornetto		**Princeton WordNet**	
H	NoS	P	NoS	P
0	833	1,26%	1	0,00%
1	62847	95,05%	93078	97,64%
2	2330	3,52%	2165	2,27%
3	98	0,15%	63	0,07%
4	11	0,02%	12	0,01%
5	2	0,00%	3	0,00%
6	0	0,00%	1	0,00%
Total	66121	100%	95323	100%

Table 5: Synset frequency and percentage of total number of synsets is shown for every number of hypernyms value in Cornetto as well as WordNet. **H** abbreviates ‘number of hypernyms’, **NoS** ‘number of synsets’ and **P** 'percentage of total number of synsets’.

Table 5 shows that Cornetto contains relatively more synsets with multiple hypernyms than Princeton WordNet. Multiple hypernyms may lead to more options to connect synsets that can be classified according to different perspectives, e.g. being both a mammal and a pet. Nevertheless, more research is needed to find a direct explanation. If these multiple hypernyms occur at the higher levels, which is often the case, they apply to large proportions of the synsets. Besides this difference, we also observe similar patterns in the correlations. In both cases, we see a significant drop in the performance of the Information Content-based measures *jcn* and *lin*. This drop in performance emphasizes the strength and weakness of these measures. Their strength is found in the fact that if the Information Content of the words is available, the correlation with human judgement can be high. However, if the Information Content is not available, which is more often the case for the larger Rubenstein & Goodenough’ gold standards, the correlation drops significantly. We do not observe this drop for the measure *res*, because this measure uses the Information Content of the least common subsumer, which is more robust than the measures *jcn* and *lin*, which are based on the Information Content of the words themselves. Finally, the differences between the scores from the WordNet::Similarity package and the WordNetTools show that we did not reproduce the results exactly. This in itself is not surprising, given the fact that Fokkens et al. (2013) showed that even replicating the results that Pedersen (2010) reports can be challenging. They showed that even if the main properties are kept stable, such as software and versions of software, variations in minor properties can lead to completely different outcomes. In addition, the reproduction learned us an interesting fact about the occassional inability of corpus statistics to distinguish between synsets. In order to use Information Content, cumulative synset frequencies are used. This creates the possibilty that a hyponym and its hypernym can have the same cumulative frequency. During our experiments, the similarity score was calculated between the synsets 'cushion#n#3' and 'pillow#n#1', where 'pillow#n#1' is a hyponym of 'cushion#n#3'. Neverthless, the cumulative frequency for both synsets is the same, which is 9. When the similarity score between these synsets was calculated for the Information Content measures, they are represented as synonyms according to these measures, which is in fact not the case in WordNet.

7 Conclusion

In this paper we described the results of re-implementing the similarity measures in a toolkit that can handle a wordnet in any language in Wordnet-LMF and the creation of a Dutch gold standard for similarity experiments similar to the English experiments. The toolkit can be tuned to handle any type of relation and thus can be used for various similarity and relatedness experiments, possibly adapted to the way the specific wordnet was built. We used these options to achieve a compatible structure to the English WordNet. We also created different variants of the Dutch gold standard to measure possible differences of interpretations of the task by the native speakers. We have shown that the Dutch gold standard is highly compatible to the English but that the Dutch wordnet performs better than the English WordNet to the same task. In the future, we will extend the toolkit to perform more operations and we will try to extend the experiment to other languages. We also want to experiment with different graphs to see the impact on the task. These graphs could reflect different degrees of relatedness depending on the relations that are selected. Such relations could also be derived from distributional properties of words and inserted into the graph, where they can be combined with wordnet relations or used separatedly.

Acknowledgements

The authors wish to thank prof dr. Jan Odijk for his help in setting up the Dutch semantic similarity experiments. Moreover, we would like to thank the members of the Computational Lexicology and Terminology Web (CLTL) @ VU University Amsterdam for their feedback. Finally, we would like to thank the participants, which were: Katelijn van Knippenberg and her students from the high school *'t Atrium* in Amersfoort, Channah van 't Wout and her students from the high school *'t Hooghe Landt* in Amersfoort, Paul Vleer and his students from the high school *Maurick College* in Vught, Renny van der Sleen and her students from the high school *RSG Trompmeesters* in Steenwijkerland, and Ina van der Wekken and her students from the high school *Jacob-Roelandslyceum* in Boxtel.

References

Christiane Fellbaum, editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eyran Ruppin (2002). *Placing search in context: The concept revisited.* In: Proceedings of the 10th international conference on World Wide Web, pages 406–414.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire (2013). *Offspring from Reproduction Problems: What Replication Failure Teaches Us.* In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Iryna Gurevych (2005). *Using the structure of a conceptual network in computing semantic relatedness.* In: Proceedings of the 2nd International Joint Conference on Natural Language Processing, Jeju Island, South Korea, pages 767–778.

Samer Hassan and Rada Mihalcea (2009). *Cross-lingual semantic relatedness using encyclopedic knowledge.* In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3, Singapore, pages 1192–1201.

Jay J. Jiang and David W. Conrath (1997). *Semantic similarity based on corpus statistics and lexical taxonomy.* In: Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X), pages 19–33.

Colette Joubarne and Diana Inkpen (2011). *Comparison of semantic similarity for different languages using the Google n-gram corpus and second- order co-occurrence measures*, Proceedings of the 24th Canadian conference on Advances in artificial intelligence, Canadian AI'11,isbn 978-3-642-21042-6 Springer-Verlag, Berlin, Heidelberg, pages 216–22.

Claudia Leacock and Martin Chodorow (1998). *Combining local context and WordNet similarity for word sense identification.* In Fellbaum, C., editor, WordNet: An electronic lexical database, MIT Press, pages 265–283.

Michael Lesk (1986). *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.* In: Proceedings of the 5th annual international conference on Systems documentation, ACM, 1986.

Dekang Lin (1998). *An information-theoretic definition of similarity.* In: Proceedings of the 15th International Conference on Machine Learning, Madison, USA, pages 296–304.

George A. Miller and Walter G. Charles (1991). *Contextual correlates of semantic similarity.* Language and Cognitive Processes, 6(1):1–28.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker (1993). *A semantic concordance.* In: Proceedings of the workshop on Human Language Technology, pages 303–308.

Nelleke Oostdijk, Martin Reynaert, Paola Monachesi, Gert-Jan Van Noord, Roeland Ordelman, Ineke Schuurman, and Vincent Vandeghinste (2008). *From D-Coi to SoNaR: a reference corpus for Dutch.* In: LREC.

Siddharth Patwardhan and Ted Pedersen (2006). *Using WordNet-based context vectors to estimate the semantic relatedness of concepts.* In: Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy, pages 1–8.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi (2004). *WordNet::Similarity: measuring the relatedness of concepts.* In: Demonstration Papers at HLT-NAACL 2004, Association for Computational Linguistics, pages 38–41.

Ted Pedersen (2010). *Information content measures of semantic similarity perform better without sense-tagged text.* In: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010), Los Angeles, USA, pages 329–332.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner (1989). *Development and application of a metric on semantic nets.* IEEE Transaction on Systems, Man, and Cybernetics, 19(1):17–30.

Philip Resnik (1995). *Using information content to evaluate semantic similarity in a taxonomy.* In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, pages 448–453.

Herbert Rubenstein and John B. Goodenough (1965). *Contextual correlates of synonymy.* Communications of the ACM, 8(10):627–633.

Piek Vossen, Attila Görög, Rubén Izquierdo, Antal van den Bosch. (2012) *DutchSemCor: Targeting the ideal sense-tagged corpus.* LREC, 584–589.

Piek Vossen, Claudia Soria, and Monica Monachini (2013). *Wordnet-LMF: a standard representation for multilingual wordnets* G. Francopoulo (ed.) LMF: Lexical Markup Framework, theory and practice, Hermes / Lavoisier / ISTE

Piek Vossen , Isa Maks, Roxane Segers, Hennie Van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten De Rijke (2013). *Cornetto: a lexical semantic database for Dutch*, P. Spyns and J. Odijk (eds): Essential Speech and Language Technology for Dutch, Results by the STEVIN-programme, Publ. Springer series Theory and Applications of Natural Language Processing, ISBN 978-3-642-30909-0.

Zhibiao Wu and Martha Palmer (1994). *Verbs semantics and lexical selection.* In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138.