What replication and localisation teach us: the case of semantic similarity measures

Marten Postma

August 2013

Supervisors: Prof. Dr. Jan Odijk & Prof Dr. Piek Vossen
University: Utrecht
Student number: 3235483

Table of Contents

1 Introduction

Many tasks in the field of Natural Language Processing make use of so-called *semantic similarity measures*, which quantify the degree to which two concepts are semantically similar. In order to know which of the semantic similarity measures is to be used for Natural Language Processing tasks, they are generally evaluated against human judgement. However, because human judgement is subjective, gold standards are created by asking a group of people to indicate the similarity of meaning of a set of word pairs. The correlation between these gold standards and the output from the semantic similarity measures gives a good indication as to which measure correlates best with human judgement.

Most research, for example Patwardhan and Pedersen (2006) and Pedersen (2010), has focused on English, using the English lexical semantic database WordNet (Miller, 1995) to compute the scores for the semantic similarity measures. The main focus of this thesis is upon getting a better understanding of the workings of semantic similarity measures by also using a different lexical semantic database in a different language, which is Cornetto (Vossen, 2006; Vossen et al., 2007, 2008) for Dutch.

In order to get a better understanding of these measures, we first inspect the previous English experiments and try to replicate them to be sure that we fully understand the process. Furthermore, we will create a Dutch gold standard and inspect the correlations between the output from the semantic similarity measures using the Dutch lexical semantic database Cornetto and the newly created Dutch gold standard.

For English, we will show that a group of semantic similarity measures approaches human judgement in a similar way. Moreover, we will stress the importance of addressing every detail of the process that leads to the results by showing that even if the main properties are kept stable, variations in minor properties can lead to completely different outcomes. Furthermore, we will present our gold standard for Dutch and how it was created. In addition, we will show that not only the properties of a semantic similarity measure determine its performance, but that the structure of the lexical semantic database also plays a crucial role.

The outline is as follows. We start by introducing the lexical semantic databases that we have used: WordNet and Cornetto, in section 2, followed by a description of the semantic similarity measures that make use of these databases in section 3. The output from these measures is compared against so-called gold standards, which are discussed in section 4. After this theoretical background, we move to a discussion on the previous English experiments that have compared the results from these measures against the gold standards. In addition, we attempt to replicate the results from these experiments in sections 5 and 6, respectively. Our own experiments are then discussed in section 7, which includes the description of the creation of the Dutch gold standard. Finally, the results and the discussion are discussed in sections 8 and 9, followed by the conclusion in section 10.

We start with a description of lexical semantic databases in the next section.

2 Lexical semantic databases

A clear understanding of lexical semantic databases is needed in order to understand the workings of semantic similarity measures. In this section, we introduce the English lexical semantic database WordNet, followed by a discussion on the semantic relations that create a structured hierarchy of concepts, also called a wordnet. We conclude with an explanation of the Dutch lexical semantic database Cornetto. We will start with a discussion on WordNet.

George Miller (Miller, 1995) was one of the first to create a semantic hierarchy of concepts, which is called WordNet. He based his wordnet on lexicalized concepts, which are represented by sets of one or more synonyms, also called *synsets.* Words are said to be synonymous if two words share at least one sense, whereas words themselves are defined as forms, i.e. strings of letters, that have a sense in a language. Words are encoded in WordNet using a 3-part name, consisting of the lemma, the part of speech and the sense number. Hence, the word ‘car.n.01’ refers to the first noun sense of the lemma *car* as listed in WordNet. In addition, synsets are often accompanied by definitions, also called *glosses.* An example of a synset is {‘car.n.01’,‘automobile.n.01’,‘machine.n.01’,‘auto.n.01’, ‘motorcar.n.01’}, in which five words represent the lexicalized concept described by the definition ‘a motor vehicle with four wheels; usually propelled by an internal combustion engine’. Synsets are linked by means of semantic relations, creating a wordnet. WordNet then can be seen as a graph in which the synsets are the nodes and the edges are formed by semantic relations.

One of the most important relations to link synsets in WordNet is the relation of hyponymy/hyperonymy. For example, the previously mentioned synset {‘car.n.01’, ‘automobile.n.01’, ‘machine.n.01’,‘auto.n.01’,‘motorcar.n.01’}, *s1*, is a hyponym of the synset {’motor vehicle.n.01’, ’automotive vehicle.n.01’}, *s2*, described by the definition ‘a self-propelled wheeled vehicle that does not run on rails’, because:

1. it is true that for all x *s2*’(x) ⇒ *s1*’(x), and

2. it is false that for all x *s1*’(x) ⇒ *s2*’(x).

The inverse of a hyponym is called a hyperonym. In addition, a hyponym can have more than one hyperonym. This relation is often also called the is-a relation, which is the term that will henceforth be used. Besides the is-a relation, WordNet also includes other semantic relations, e.g. antonymy, entailment and meronymy.

The Dutch lexical semantic database Cornetto (Vossen, 2006; Vossen et al., 2007, 2008) contains the *Referentiebestand Nederlands* (Martin and Maks, 2005) and the *Dutch WordNet* (Vossen, 1998). Cornetto was based on the English Wordnet and is very similar in its design. However, there are differences between WordNet and Cornetto. WordNet 3.0 contains 117.659 unique synsets, whereas Cornetto contains approximately 70.000 synsets. In addition, the design of the top levels in Cornetto differs from that in WordNet. The top levels of Cornetto are quite general, whereas those of WordNet are more differentiated.

3 Semantic similarity measures

The is-a hierarchy, as well as other semantic relations inside lexical semantic databases, can be used in quantifying the degree to which two synsets are semantically similar. This quantifying of semantic similarity is done with *semantic similarity measures* and is widely used inside the field of Natural Language Processing. For example, they are used to find *malapropisms* (Hirst and St-Onge, 1998) and in *word sense disambiguation* (Banerjee and Pedersen, 2003). Even outside of the field of NLP, they are used in the *biomedical domain* (Pedersen et al., 2007). There are two main types of semantic similarity measures:

1. Semantic similarity
2. Semantic relatedness

We follow Strube and Ponzette (Strube and Ponzetto, 2006) in stating that semantic similarity only takes into account the is-a relation, whereas semantic relatedness can take into account all semantic relations inside the database, which includes the is-a relation. We will start with a discussion on the concepts *path length*, *LCS* and *range*, which are crucial to most semantic measures. This is followed by a discussion on the following semantic measures: Path, Wu & Palmer, Leacock & Chodorow, Resnik, Jiang & Conrath, Lin, Hirst and St-Onge, Adapted Lesk, Vector, and Vector pairs. Finally, we will discuss which semantic similarity measures we implement in our Dutch experiments using Cornetto.

It is clear that a lexical semantic database can be seen as a graph in which the synsets are the nodes and the edges, i.e. semantic relations, connect the nodes. Using these elements, it is possible to create a path from one synset *s1* to another synset *s2*. In general, there are two main ways to calculate the length of the path from *s1* to *s2*. This can either be done by counting the number of edges that separate *s1* and *s2* or the number of nodes. When the path length of the same synsets is calculated, we get the minimum path length, which is 0 when we use edge counting. However, using node counting, this same path has a length of 1. Due to the fact that the path length acts as the denominator in inverse functions, it is not preferable that the path length can be 0, which is why node counting is most often used and hence also in this thesis. In addition, it is also possible that multiple paths connect two nodes, because of the fact that a hyponym can have multiple hyperonyms. Whenever this occurs, the path length is always the shortest path connecting the nodes. The shortest path length, or $dist_{\mathrm{node}}(s_1, s_2)$, can then be defined as the shortest path length between synsets *s1* and *s2* using node counting. In addition, the term *LCS* is often used in semantic similarity measures. This abbreviation stands for the least common subsumer, where for synsets *s1* and *s2* LCS(s1,s2) is defined as the deepest hyperonym in the lexical semantic database that both synsets share. Finally, many of the similarity measures that will be discussed have a different *range* of possible output values. For some, the range of values is between 0 and 1, whereas others can have much higher output scores. This might raise the question whether it is possible to compare the results from the similarity measures. However, we avoid this problem by using a correlation that is based on ranks instead of on absolute values. This correlation measure will be discussed in more detail in subsection 7.7. After discussing these important terms, we will now discuss each of the semantic similarity measures.

The similarity score for **path** (Rada et al., 1989) is calculated taking the multiplicative inverse of $dist_{\text{node}}(s_1, s_2)$. The maximum occurs when two synsets are the same, the score is then 1.0. The minimum approaches zero as the path length increases. This is dependent on the maximum width and depth of the lexical semantic database. Intuitively, the shorter the path, the more similar two synsets are. The formula can be found in equation 1:

$$Sim_{\text{path}}(s1, s2) = \frac{1}{dist_{\text{node}}(s_1, s_2)} \tag{1}$$

In order to refer to this similarity measure, we will be using the term *path*.

Wu & Palmer (Wu and Palmer, 1994) added the notion of depth into their semantic measure. We define the depth of a synset s as follows:

$$depth(s) = dist_{\text{node}}(s, r) \tag{2}$$

where r is the root. The minimum depth using node counting is 1. Next, the depth of the LCS(s1,s2) is calculated. Furthermore, the path length between each synset and the LCS(s1,s2) is calculated. The closer the depth of the LCS(s1,s2) is to the depth of the two synsets, the higher the semantic similarity score will be. In addition, the deeper the synsets are in the lexical semantic database, the higher the score will be. Wu & Palmer originally defined the following formula, which can be found in equation 3:

$$Sim_{\text{OriginalWuPalmer}}(s1, s2) = \frac{2 * depth(LCS(s_1, s_2))}{\begin{array}{r} dist_{\text{node}}(s_1, LCS(s_1, s_2)) + dist_{\text{node}}(s_2, LCS(s_1, s_2)) \\ + 2 * depth(LCS(s_1, s_2)) \end{array}} \tag{3}$$

Resnik (Resnik, 1999) adapted this formula slightly. The formula can be found in equation 4:

$$Sim_{\text{wupalmer}}(s1, s2) = \frac{2 * depth_{\text{node}}(LCS(s1, s2))}{depth_{\text{node}}(s_1) + depth_{\text{node}}(s2)} \tag{4}$$

We will use the adapted formula by Resnik in this thesis. The term to refer to this semantic measure is *wup*.

Besides the path length, the following measure, designed by **Leacock & Chodorow** (Leacock and Chodorow, 1998), also takes into account the depth of the lexical semantic database in which the synsets are situated, which is abbreviated by the letter D in the formula. More specifically, D denotes the maximum depth of the lexical semantic database. This is done in order to be able to compare scores from this measure across different lexical semantic databases. The negative log-likelihood function is probably used for a practical reason to make sure that a higher score also corresponds to a higher similarity, although this is not mentioned in the original paper. The formula can be found in equation 5:

$$Sim_{\text{lch}}(s1, s2) = -log(\frac{dist_{\text{node}}(s_1, s_2)}{2 * D}) \tag{5}$$

This semantic similarity measure will be referred to by the term *lch*.

Resnik (Resnik, 1995) is the first to use the notion of Information Content in a semantic similarity measure. By doing this, Resnik moves away from the measures that solely rely on the distance between two synsets in a lexical semantic database. The main reason for doing this is that he claims that distance measures rely on the idea that the edges between nodes represent uniform distances, which he claims is not the case in most lexical semantic databases. Instead, he bases his measure on the estimated probability of a synset. The lower the probability of a synset, p(s) , the higher the Information Content. The main idea behind this is that Resnik claims that abstract synsets carry less information than more specific synsets. In order to also show this quantitatively, the log-likelihood function is used. The Information Content of a synset, IC(s), can then be defined as taking the negative log of the estimated probability of a synset: -log(p(s)). In order to obtain the estimated probabilities of synsets, Resnik presents two ways. Firstly, the estimated probabilities can be used from a sense-tagged corpus. For English, the sense-tagged corpus SemCor (Miller et al., 1993) is available. Secondly, the estimated probabilities can be obtained by using the lemma frequencies from a corpus. Each synset that is associated with a word lemma receives an equal share of the lemma frequency. For example, if two synsets are associated with the word lemma *example*, then each of these synsets will receive a count of 0.5 for each occurence of this lemma in the corpus. Not only the synsets themselves receive this count, but also all hyperonyms of these synsets. Resnik obtains good results using both ways to estimate probabilities of synsets. More recently, Pedersen (2010) obtained good results using the second way to estimate the probabilities of the synsets, using a corpus of 1.2 million tokens. The results improved as the corpus size increased. In order to calculate the similarity between two synsets, the Information Content is taken of the least common subsumer. When no estimate of the probability of a synset is available, a default score is generally assigned, for example 0 or -1.0. For Dutch, there was no equivalent to SemCor available, which is why we used the frequencies of the lemmas in the Dutch corpus called SoNaR (Oostdijk et al., 2008) The formula can be found in equation 6:

$$Sim_{\text{res}}(s1, s2) = IC(LCS(s_1, s_2)) \tag{6}$$

In order to refer to this semantic measure, the term *res* will be used.

In addition to the Information Content of the LCS, **Jiang and Conrath** (Jiang and Conrath, 1997) also use the Information Content of the synsets themselves. The closer the Information Content of LCS is to IC(s1) and IC(s2), the lower the score will be. Again, when no estimate of the probability of a synset is available, a default score is assigned. The original formula can be found in equation 7:

$$Sim_{\text{jcn}}(s1, s2) = IC(s_1) + IC(s_2) - 2 * IC(LCS(s_1, s_2)) \tag{7}$$

In the package that we use to calculate the scores of the semantic similarity measures in WordNet, WordNet::Similarity (Pedersen et al., 2004), the multiplicative inverse is taken of the result from the original formula of Jiang and Conrath. Once again, this is done to make sure that a higher score also represents a higher similarity. One disadvantage of this approach is that if IC(s1) + IC(s2) = IC(LCS(s1,s2)), a score of 1 would be divided by 0, which is undefined. Hence, whenever this occurs, the smallest possible distance greater than zero is chosen and the multiplicative inverse of that distance is returned. We will refer to this semantic measure with the term *jcn*.

Lin (Lin, 1998) slightly adapts the formula by Jiang and Conrath. Again, a default value is returned if no estimate of the probability of one of the synsets is available. The formula can be found in equation 8:

$$Sim_{\text{jcn}}(s1, s2) = \frac{2 * IC(LCS(s_1, s_2))}{IC(s_1) + IC(s2)} \tag{8}$$

We will refer to this semantic measure with the term *lin*.

The explanation of the similarity measure *lin* ends the discussion on similarity measures. In the previous English experiments in which semantic similarity measure are discussed, four more measures are used, which are relatedness measures. Since the scope of this thesis is on similarity and not on relatedness, we will not discuss them extensively. **Hirst and St-Onge** (Hirst and St-Onge, 1998) created a semantic relatedness measure, which was based on so-called **lexical chains**. This measure takes into account all semantic relations, not only the is-a relation. Using these relations, **lexical chains** are formed, which are all paths between synsets *s1* and *s2* using all semantic relations in the lexical semantic database. The shorter the paths and the less the path directions change, the higher the score will be. We will refer to this semantic measure with the term *hso*. Finally, three semantic relatedness measures are not primarily based on the is-a hierarchy, but on the overlap in the definitions of the synsets, also called *glosses*. These measures are *Adapted Lesk* (Banerjee and Pedersen, 2003), *Gloss Vector* (Patwardhan and Pedersen, 2006) and *Vector pairs* (Patwardhan and Pedersen, 2006).

We have seen that many semantic similarity measures require different elements in order to work. Some only need an is-a hierarchy, such as *path*, whereas others, such as *hso*, require all semantic relations. However, it was predominantly discussed how each of them works in WordNet. In order to implement these measures in a different lexical semantic database, for example Cornetto, we need to be sure that all elements that are available for WordNet, are also available for Cornetto. One of the elements that is essential for the measures *Adapted Lesk*, *Gloss Vector*, and *Vector pairs* is that the glosses in Cornetto are of a certain quality and length. However, due to the fact that this is not the case, we will not use them in this thesis. We will not implement any of the relatedness measures, hence we will also not use *hso*. In addition, the element that is mostly used for the measures that rely on Information Content, namely a sense-tagged corpus, is not available for Dutch. We approach this by using the frequencies of the lemmas in the Dutch corpus called SoNaR (Oostdijk et al., 2008). Finally, there are no problems implementing the measures *path*, *lch*, and *wup*. In summary, we have implemented the measures *path*, *lch*, *wup*, *res*, *jcn*, and *lin*.

4 Gold standards

In order to determine how well similarity measures approach human judgement, a gold standard is needed. However, human judgement is highly subjective. That is why most studies use two gold standards developed by Rubenstein & Goodenough (Rubenstein and Goodenough, 1965) and Miller & Charles (Miller and Charles, 1991). We will first discuss these original gold standards, followed by a discussion on the WordSimilarity-353 Test Collection (Finkelstein et al., 2002). Finally, we will discuss two studies in which the original gold standards were translated.

Rubenstein & Goodenough created a list of 65 word pairs of which they claim that it ranges (Rubenstein and Goodenough, 1965, p. 627):

> '*from highly synonymous pairs to semantically unrelated pairs*'

The participants were given 65 slips of paper. One word pair was printed on one slip of paper. Participants were then asked to first order the slips of paper they had received on similarity of meaning and then assign a value from 0 to 4 to each word pair. In total, 51 participants assigned values to the word pairs. These 51 participants were divided into two groups. The first group, containing 15 subjects, first gave synonymy judgements on 48 pairs, of which 36 pairs were later selected for the study. In the second session of this group, they gave synonymy judgements on the 65 pairs finally selected. In addition, 36 different subjects also rated the 65 pairs finally selected. The mean ratings for each word pair can be found in appendix A. We would like to mention three points of criticism against the methodology used in this experiment, concerning the limited number of unique words used in the gold standard, the lack of information as to which meaning of a word participants had in mind, and the uncertainty as to what is meant by *similarity of meaning*. Firstly, only 48 nouns were used to create the word pairs, which means that some word forms occur five times in the list and some just once. Secondly, we do not know which sense of the word forms participants had in mind, which will be important when we will be translating these datasets. For example, when seeing the word form *string*, we are not sure that the participant had the meaning of ‘piece of rope’ in mind instead of ‘a piece of womens underwear’. One way to get an idea of which meaning participants had in mind is to look at the scores from human judgement. For example, WordNet lists two senses for the noun *asylum*:

1. ‘a shelter from danger or hardship’
2. ‘a hospital for mentally incompetent or unbalanced person’

In the results from Miller & Charles and Rubenstein & Goodenough, we observe that the correlation with *madhouse* is very high. Hence, we believe that participants had the second sense as listed for *asylum* in WordNet in mind. Finally, participants were asked to rate on similarity of meaning. However, it is not completely clear what is meant by this term.

Miller & Charles (Miller and Charles, 1991) used a subset of 30 word pairs from the study by Rubenstein & Goodenough. They used 10 word pairs from word pairs that had received a score from 3 to 4, 10 word pairs that had received a score from 1 to 3 and 10 word pairs that had received a score from 0 to 1. The 38 participants were asked to rate each word on a five-point scale from 0 to 4. The results can be found in appendix B. The same points of criticism apply to this study.

The WordSimilarity-353 Test Collection (Finkelstein et al., 2002) contains two sets of English word pairs with similarity scores assigned by humans. These datasets differ with respect to the previous ones concerning their size and the instructions that were given to the participants. 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. 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.

A number of studies has 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. Finally, a sixth person evaluated the translation. They then asked 5 participants to rate the Spanish word pairs. Because of the fact that the correlation with the original datasets was 0.86. only one translator translated the datasets into Arabic and Romanian. Finally, Gurevych (Gurevych, 2005) translated the datasets into German. However, no instructions, as to how it was done, were provided.

5 Previous English experiments

In order to establish which semantic similarity measures approach human judgement the best using WordNet, Patwardhan & Pedersen (Patwardhan and Pedersen, 2006) and Pedersen (Pedersen, 2010) conducted a number of experiments. We will first discuss the experiment by Patwardhan & Pedersen followed by a discussion on the experiment by Pedersen.

Patwardhan & Pedersen introduce the similarity measure called *Gloss Vector* in this paper. This similarity measure has been mentioned in section 3. In order to evaluate this new similarity measure, they compare the results from existing similarity measures against the results from the similarity measure *Gloss Vector*. We will discuss the (1) material, (2) similarity measures, (3) procedure, and (4) results from this experiment.

(1) **Material**: The human judgements scores are those from the study by Rubenstein & Goodenough and by Miller & Charles. The package to calculate the similarity measures scores is called WordNet::Similarity (Pedersen et al., 2004). The modules used to calculate the similarity measures for this study are:

1. WordNet version 2.1
2. WordNet-QueryData version 1.39
3. Text-Similarity version 0.02
4. WordNet-Similarity version 1.02

(2) **Similarity measures**: The six similarity measures used in this study are *lch*, *res*, *jcn*, *lin*, *Adapted Lesk*, and *Gloss Vector*. In order to calculate the relatedness score between two words, for example *tree* and *shrub*, the scores between all senses of *tree* and *shrub* are calculated. The highest score is then chosen.

(3) **Procedure**: The Spearman's rank correlation coefficient ρ (Spearman, 1904) is used to evaluate the correlation between the values created by similarity measures and by human judgement. This correlation works with rankings of the two sets. If the rankings are exactly the same, the score is 1.0. A completely reversed ranking obtains a score of -1.0.

(4) **Results**: The results can be found in table 1:

Table 1: Correlation with the datasets by Miller & Charles and Rubenstein & Goodenough for six similarity measures.

Measure	**MC**	**RG**
Gloss Vector	0.91	0.90
Adapted Lesk	0.81	0.83
Jcn	0.73	0.75
Res	0.72	0.72
Lin	0.70	0.72
Lch	0.74	0.70

Following the results from table 1, it is clear that the relatedness measures *Gloss Vector* and *Adapted Lesk* seem to correlate best with human judgement using WordNet. The similarity measures all seem to perform equally well between 0.70 and 0.75.

Pedersen (Pedersen, 2010) replicated the results from Patwardhan & Pedersen. He also added two similarity measures. These are *path* and *wup*. The versions of the modules were changed in the methodology. Firstly, version 2.05 of WordNet-Similarity was used and version 3.0 of WordNet and SemCor. Wordnet-QueryData version 1.49 was used and Text-Similarity version 0.08. The results can be found in table 2.

Table 2: Correlation with the datasets by Miller & Charles and Rubenstein & Goodenough for eight similarity measures.

Measure	**MC**	**RG**
Gloss Vector	0.89	0.73
Adapted Lesk	0.83	0.68
Wup	0.74	0.69
Lch	0.71	0.70
Path	0.68	0.69
Jcn	0.72	0.51
Lin	0.73	0.58
Res	0.74	0.60

Once again, we observe that the relatedness measures seem to outperform the similarity measures. The scores of the similarity measures seem to be around 0.70, whereas the measures that are based on Information Content do not seem to correlate well with the dataset by Rubenstein & Goodenough.

6 Replication experiments

We have tried to replicate the experiments of the previous section in order to be sure that we fully understand the process. We will start with a discussion on the first replication attemps, followed by the methodology and the first observations. In addition, we show what the possible variations for the results can be if not sufficient information is available to replicate the results. [1]

6.1 Replication attempts

First, an attempt was made to reproduce the results reported in (Patwardhan and Pedersen, 2006) and (Pedersen, 2010) on the English WordNet using their WordNet::Similarity web-interface. [2] Results differed from those reported in the aforementioned works, even when using the same versions as the original, WordNet::Similarity-1.02 and WordNet 2.1 (Patwardhan and Pedersen, 2006) and WordNet::Similarity-2.05 and WordNet 3.0 (Pedersen, 2010), respectively. [3]

The fact that results of similarity measures on WordNet can differ even while the same software and same versions are used indicates that properties which are not addressed in the literature may influence the output of similarity measures. We therefore conducted a range of experiments that, in addition to searching for the right settings to replicate results of previous research, address the following questions:

1. Which properties have an impact on the performance of WordNet similarity measures?
2. How much does the performance of individual measures vary?
3. How do commonly used measures compare when the variation of their performance are taken into account?

6.2 Methodology and first observations

The questions above were addressed in two stages. In the first stage, Fokkens, who was not involved in the first replication attempt implemented a script to calculate similarity measures using Word- Net::Similarity. This included the following similarity measures: *wup*, *lch*, *res*, *res*, *jcn*, *lin*, *Adapted Lesk*, *hso*, *Gloss Vector* and *Vector pairs.*

[1] The following section is the result of joint research alongside with Antske Fokkens, Marieke van Erp, Ted Pedersen, Piek Vossen and Nuno Freire. The title of the paper is ‘Offspring from Reproduction Problems: What Replication Failure Teaches Us’. It has been presented as the 51st Annual Meeting of the Association for Computational Linguistics in Sofia, Bulgaria, August 4-9 2013, where it was nominated for the best paper award. It finished runner-up. The paper is currently in preprint, which can be found in appendix L.

[2] Obtained from http://talisker.d.umn.edu/cgi-bin/similarity/similarity.cgi, WordNet::Similarity version 2.05. This web interface has now moved to http://maraca.d.umn.edu.

[3] Wordnet::Similarity was obtained from http://search.cpan.org/dist/WordNet-Similarity.

Consequently, settings and properties were changed systematically and shared with Pedersen who attempted to produce the new results with his own implementations. First, we made sure that the script implemented by Fokkens could produce the same WordNet similarity scores for each individual word pair as those used to calculate the ranking on the mc-set by Pedersen (2010). Finally, the gold standard and exact implementation of the Spearman ranking coefficient were compared.

Differences in results turned out to be related to variations in the **experimental setup**. First, we made different assumptions on the restriction of part-of-speech tags (henceforth 'PoS-tag') considered in the comparison. Miller and Charles (Miller and Charles, 1991) do not discuss how they deal with words with more than one PoS-tag in their study. Pedersen therefore included all senses with any PoStag in his study. The first replication attempt had restricted PoS-tags to nouns based on the idea that most items are nouns and subjects would be primed to primarily think of the noun senses. Both assumptions are reasonable. Pos-tags were not restricted in the second replication attempt, but because of a bug in the code only the first identified PoS-tag ('noun' in all cases) was considered. We therefore mistakenly assumed that PoS-tag restrictions did not matter until we compared individual scores between Pedersen and the replication attempts.

Second, there are two gold standards for the Miller and Charles set: one has the scores assigned during the original experiment run by Rubenstein and Goodenough , the other has the scores assigned during Miller and Charles's own experiment. The ranking correlation between the two sets is high, but they are not identical. Again, there is no reason why one gold standard would be a better choice than the other, but in order to replicate results, it must be known which of the two was used. Third, results changed because of differences in the treatment of ties while calculating Spearman ρ. The influence of the exact gold standard and calculation of Spearman ρ could only be found because Pedersen could provide the output of the similarity measures he used to calculate the coefficient. It is unlikely we would have been able to replicate his results at all without the output of this intermediate step. Finally, results for *lch*, *Adapted Lesk* and *wup* changed according to measure specific configuration settings such as including a PoS-tag specific root node or turning on normalisation.

In the second stage of this research, we ran experiments that systematically manipulate the influential factors described above. In this experiment, we included both the mc-set and the complete rgset. The implementation of Spearman ρ used in (Pedersen, 2010) assigned the lowest number in ranking to ties rather than the mean, resulting in an unjustified drop in results for scores that lead to many ties. We therefore experimented with a different correlation measure, Kendall tau coefficient τ (Kendall, 1938) rather than two versions of Spearman ρ.

6.3 Variation per measure

All measures varied in their performance. The complete outcome of our experiments (both the similarity measures assigned to each pair as well as the output of the ranking coefficients) are included in the data set provided at http://github.com/antske/WordNetSimilarity. Table 3 presents an overview of the main point we wish to make through this experiment: the minimal and maximal results according to both ranking coefficients.

Table 3: Variation WordNet measures results

Measure	Spearman ρ		Kendall τ		ranking variation
	min	max	min	max	
path based similarity					
Path	0.70	0.78	0.55	0.62	1-8
Wup	0.70	0.79	0.53	0.61	1-6
Lch	0.70	0.78	0.55	0.62	1-7
path based information content					
Res	0.65	0.75	0.26	0.57	1-8
Lin	0.49	0.73	0.36	0.53	6-10
Jcn	0.46	0.73	0.32	0.55	5,7-11
path based relatedness					
Hso	0.73	0.80	0.36	0.41	1-3,5-10
dictionary and corpus based relatedness					
Vector pairs	0.40	0.70	0.26	0.50	7-11
Gloss vector	0.48	0.92	0.33	0.76	1,2,4,6-11
Adapted Lesk	0.66	0.83	-0.02	0.61	1-8,11,11

Results for similarity measures varied from 0.06- 0.42 points for Spearman ρ and from 0.05-0.60 points for Kendall τ. The last column indicates the variation of performance of a measure compared to the other measures, where 1 is the best performing measure and 12 is the worst. [4] For instance, *path* has been best performing measure, second best, eighth best and all positions in between, vector has ranked first, second and fourth, but also occupied all positions from six to eleven.

In principle, it is to be expected that numbers are not exactly the same while evaluating against a different data set (the mc-set versus the rg-set), taking a different set of synsets to evaluate on (changing PoS-tag restrictions) or changing configuration settings that influence the similarity score. However, a variation of up to 0.44 points in Spearman ρ and 0.60 in Kendall τ [5] leads to the question of how indicative these results really are. A more serious problem is the fact that the comparative performance of individual measures changes. Which

[4] Some measures ranked differently as their individual configuration settings changed. In these cases, the measure was included in the overall ranking multiple times, which is why there are more ranking positions than measures.

[5] Subsection 6.4 explains why the variation in Kendall is this extreme and ρ is more appropriate for this task.

measure performs best depends on the evaluation set, ranking coefficient, PoS-tag restrictions and configuration settings. This means that the answer to the question of which similarity measure is best to mimic human similarity scores depends on aspects that are often not even mentioned, let alone systematically compared.

6.4 Variation per category

For each influential category of experimental variation, we compared the variation in Spearman ρ and Kendall τ, while similarity measure and other influential categories were kept stable. The categories we varied include WordNet and WordNet::Similarity version, the gold standard used to evaluate, restrictions on PoS-tags, and measure specific configurations. Table 4 presents the maximum variation found across measures for each category. The last column indicates how often the ranking of a specific measure changed as the category changed, e.g. did the measure ranking third using specific configurations, PoS-tag restrictions and a specific gold standard using WordNet 2.1 still rank third when WordNet 3.0 was used instead? The number in parentheses next to the different ranks in the table presents the total number of scores investigated. Note that this number changes for each category, because we compared two WordNet versions (WN version), three gold standard and PoS-tag restriction variations and configuration only for the subset of scores where configuration matters.

Table 4: Variation per category

Variation	Maximum difference		Different
	Spearman ρ	Kendall τ	rank (tot)
WN version	0.44	0.42	223 (252)
gold standard	0.24	0.21	359 (504)
PoS-tag	0.09	0.08	208 (504)
configuration	0.08	0.60	37 (90)

There are no definite statements to make as to which version (Patwardhan and Pedersen, 2006; Pedersen, 2010) , PoS-tag restriction or configuration gives the best results. Likewise, while most measures do better on the smaller data set, some achieve their highest results on the full set. This is partially due to the fact that ranking coefficients are sensitive to outliers. In several cases where PoS-tag restrictions led to different results, only one pair received a different score. For instance, *path* assigns a relatively high score to the pair chord-smile when verbs are included, because the hierarchy of verbs in WordNet is relatively flat. This effect is not observed in *wup* and *lch* which correct for the depth of the hierarchy. On the other hand, *res*, *lin* and *jcn* score better on the same set when verbs are considered, because they cannot detect any relatedness for the pair crane-implement when restricted to nouns.

On top of the variations presented above, we notice a discrepancy between the two coefficients. Kendall τ generally leads to lower coefficiency scores than

Spearman ρ. Moreover, they each give different relative indications: where Adapted Lesk achieves its highest Spearman ρ, it has an extremely low Kendall τ of 0.01. Spearman ρ uses the difference in rank as its basis to calculate a correlation, where Kendall τ uses the number of items with the correct rank. The low Kendall τ for Adapted Lesk is the result of three pairs receiving a score that is too high. Other pairs that get a relatively accurate score are pushed one place down in rank. Because only items that receive the exact same rank help to increase τ, such a shift can result in a drastic drop in the coefficient. In our opinion, Spearman ρ is therefore preferable over Kendall τ. We included τ , because many authors do not mention the ranking coefficient they use (cf. (Budanitsky and Hirst, 2006), (Resnik, 1995)) and both ρ and τ are commonly used coefficients.

Except for WordNet, which (Budanitsky and Hirst, 2006) hold accountable for minor variations in a footnote, the influential categories we investigated in this paper, to our knowledge, have not yet been addressed in the literature. Cramer (2008) points out that results from WordNet-Human similarity correlations lead to scattered results reporting variations similar to ours, but she compares studies using different measures, data and experimental setup. This study shows that even if the main properties are kept stable, results vary enough to change the identity of the measure that yields the best performance. Table 3 reveals a wide variation in ranking relative to alternative approaches. Results in Table 4 show that it is common for the ranking of a score to change due to variations that are not at the core of the method.

This study shows that it is far from clear how different WordNet similarity measures relate to each other. In fact, we do not know how we can obtain the best results. This is particularly challenging, because the best results may depend on the intended use of the similarity scores (Meng et al., 2013). This is also the reason why we presented the maximum variation observed, rather than the average or typical variation (mostly below 0.10 points). The experiments presented in this paper resulted in a vast amount of data. An elaborate analysis of this data is needed to get a better understanding of how measures work and why results vary to such an extent. We leave this investigation to future work. If there is one takehome message from this experiment, it is that one should experiment with parameters such as restrictions on PoS-tags or configurations and determine which score to use depending on what it is used for, rather than picking something that did best in a study using different data for a different task and may have used a different version of WordNet.

7 Experiments

After succesfully replicating the experiments in the last section, we are confident that we understand the process that led to the results. Hence, we can continue with the second major step in this thesis, which is to also run the experiments for Dutch. The design of these experiments will be described in this section. We start with a discussion about the resources that were used, followed by a discussion on how the English datasets were localised. Furthermore, the design of the experiments will be discussed, followed by a description of the participants. Finally, the procedure for the participants as well as the similarity measures are discussed, followed by the analysis.

7.1 Resources

Three datasets of English word pairs are used. The datasets from Rubenstein & Goodenough (Rubenstein and Goodenough, 1965) and Miller & Charles (Miller and Charles, 1991) are localised into Dutch. The original English datasets can be found in appendices A and B, respectively. The dataset WordSim353 Test Collection (Finkelstein et al., 2002) is used to provide example word pairs in the instructions. In order to calculate relative frequencies from the English nouns, the English sense-tagged corpus SemCor (Miller et al., 1993) is used. [6] For Dutch, the frequencies of the lemmas in the Dutch corpus called SoNaR (Oostdijk et al., 2008) are used. WordNet (Miller, 1995) is used for determining the correct sense of a particular word form. [7]

7.2 Dutch localisation

In this subsection, the design of the localisation of the gold standards by Rubenstein & Goodenough and Miller & Charles will be discussed. We will start with a short discussion on the term *localisation* and the words in the datasets, followed by a discussion on the localisation itself.

We opted for the term *localisation* instead of the term *translation*, because of a practical reason. We did not only want to translate the datasets, but we also wanted to make them comprehensible for the participants. A good example of this process is the word *mound* in the datasets. A native speaker of English will probably know the meaning of 'the slight elevation on which the pitcher stands' of this word. However, this meaning is culturally dependent. Speakers of Dutch will not all know this meaning, which is why we opted for the well-known meaning of 'natural mound'. Since we did not only translate the term, but made sure that the participants understood the terms, we call the process *localisation* instead of *translation.* Because the words used by Miller & Charles are a subset of the words used by Rubenstein & Goodenough, and because nouns are used more than once in both experiments, there are only 49 unique nouns used in both experiments. There are 49 unique nouns instead of the previously mentioned 48 unique nouns, because Miller & Charles made one change to the dataset by Rubenstein & Goodenough. Whenever Rubenstein & Goodenough used the word *cord*, Miller & Charles use the word *chord.*

Inspired by Hassan and Mihalcea (2009) as discussed in section 4 , The following general procedure is followed in the localisation of the 49 nouns:

[6] The version of Semcor is used that is associated with WordNet 3.0.

[7] Wordnet version 3.0 is used.

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

2. Following the results in 1, a Dutch localisation is chosen for each noun.

3. In addition, it is checked whether the relative frequency of the Dutch and the English nouns are in the same class of relative frequency. This is done in order to make sure that there are no outliers. A localisation 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. In appendix C, the word sense disambiguation for the set of 49 English nouns is shown. For clarity, we will repeat the example already described in section 4. For example, WordNet lists two senses for the noun *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 nouns.

The next step is to localise all English nouns into Dutch. In appendix D, the localisation with their explanations are shown. 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 are closely related to the English synonyms. For example, due to the fact that there are not two common Dutch synonyms for *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 frequency of the English noun and its localisation was checked. the English sense-tagged corpus SemCor (Miller et al., 1993) was used to calculate relative frequencies from the English nouns. For Dutch, 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 noun and its Dutch counterpart were located in the same class of relative frequency. The results can be found in appendix E. 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.

7.3 Design experiments

Using the localisation as described in subsection 7.2, our main goal is to reproduce the experiments by Miller & Charles and Rubenstein & Goodenough for Dutch. However, as discussed in section 4, the explanation of the concept *Similarity of meaning* in those experiments was not completely clear. This is why we opted to reproduce each experiment with three different kinds of instructions as to what is meant by *Similarity of meaning.* These three kinds of instructions are *Relatedness*, *Similarity*, and *No explanation.* This results in a total of six experiments, since each dataset is conducted with all three kinds of instructions. We will first explain how each kind of instruction concerning *Similarity of meaning* is explained in the experiments, followed by a discussion of the six experiments.

The three kinds of explanations for *Similarity of meaning* are: **Similarity**, **Relatedness** and **No Instruction**.

Relatedness

Participants were asked to judge the similarity of meaning of each word pair on a five-point scale from 0 to 4. Instructions were given as to what was meant by the term *Similarity of meaning.* In this variant, the instructions were focused towards *relatedness.* An example was given for each value that could be assigned to a word pair.

I Value 0: *komkommer* ‘cucumber’ & *professor* ‘professor’. Explanation: it is very unlikely that *komkommer* & *professor* occur in the same situation. The word pair was selected for inclusion in the instruction, because the value of the word pair ‘cucumber’ & ‘professor’ in the dataset WordSim353 is 0.31 on a scale from 0 tot 10.

II Value 1: *probleem* ‘problem’ & *vliegveld* ‘airport’. Explanation: it is more likely that *probleem* & *vliegveld* occur together than for value 0, because there are sometimes problems at airports. The word pair was selected for inclusion in the instruction, because the value of the word pair ‘problem’ & ‘airport’ in the dataset WordSim353 is 2.38 on a scale from 0 to 10.

III Value 2: *auto* ‘car’ & *vlucht* ‘flight’. Explanation: The ‘similarity of meaning’ of *auto* ‘car’ & *vlucht* ‘flight’ is higher than for value 1, because one can escape in a car. The word pair was selected for inclusion in the instruction, because the value of the word pair ‘car’ & ‘flight’ in the dataset WordSim353 is 4.94 on a scale from 0 to 10.

IV Value 3: *computer* ‘computer’ & *internet* ‘internet’. Explanation: these words have a high ‘similarity of meaning’, because it’s very likely that *computer* 'computer’ & *internet* ‘internet’ occur in the same situation. The word pair was selected for inclusion in the instruction, because the value of the word pair ‘computer’ & ‘internet’ in the dataset WordSim353 is 7.58 on a scale from 0 to 10.

V Value 4: *fiets* 'bicycle' & *rijwiel* 'bike'. Explanation: these words are synonyms, which makes that the highest similarity of meaning is assigned. The word pair is not in the dataset WordSim353. The word pair was chosen, because it consists of synonyms.

Similarity

Participants were asked to judge the similarity of meaning of each word pair on a five-point scale from 0 to 4. Instructions were given as to what was meant by the term *Similarity of meaning.* In this variant, the instructions were focused towards *similarity.* An example was given for each value that could be assigned to a word pair.

I Value 4: *fiets* 'bicycle' & *rijwiel* 'bike'. Explanation: value 4 is found when one of the two words can be used instead of the other, and the other way around. *fiets* 'bicycle' can be used instead of *rijwiel* 'bike' and the other way around. This word pair was chosen, because the two words are synonyms.

II Value 3: *aardappelmesje* 'potato peeler' & *mes* 'knife'. Explanation: value 3 is found when one of the two words can be used instead of the other, but not the other way around. *mes* 'knife' can be used instead of *aardappelmesje* 'potato peeler', but not the other way around. This word pair was chosen, because *aardappelmesje* 'potato peeler' is a hyponym of *mes* 'knife'.

III Value 2: example 1: *aardappelmesje* 'potato peeler' & *eetgerei* 'tableware' ; example 2: *vliegtuig* 'airplane' & *auto* 'car'. Explanation: value 2 is found if one of the two words can be used instead of the other, but not the other way around and it is somewhat more vague. *eetgerei* 'tableware' can be used instead of *aardappelmesje* 'potato peeler', but not the other way around. This word pair was chosen, because *eetgerei* 'tableware' is the hyperonym of the hyperonym of *aardappelmesje* 'potato peeler'. *vliegtuig* 'airplane' & *auto* 'car' can both be used instead of *vervoersmiddel* 'means of transport'.

IV Value 1: example 1: *aardappelmesje* 'potato peeler' & *materiaal* 'material' ; example 2: *mens* 'human' & *kat* 'cat'. Explanation: Value 1 is assigned if one of the two words can be used instead of the other, but not the other way around and it's even more general. *materiaal* 'material' can be used instead of *aardappelmesje* 'potato peeler', but not the other way around. This word pair was chosen, because the path distance between these two words is 5. *mens* 'human' & *kat* 'cat' can both be used as *levende wezens* 'living creatures'.

V Value 0: example 1: *aardappelmesje* 'potato peeler' & *iets* 'something' ; example 2: *komkommer* 'cucumber' & *professor* 'professor'. Explanation: Value 0 is found if one of the two words cannot be used instead of the other or if the relation is too general. *iets* 'something' can be used instead of *aardappelmesje* 'potato peeler', but it is too general. This word pair was chosen, because the path distance between these two words is 7. In addition, the relation between *komkommer* 'cucumber' & *professor* 'professor' is too general.

No explanation
Participants were asked to judge the similarity of meaning of each word pair. The only instructions that were given were that the highest score is assigned if two words are synonyms.

Once the instructions concerning *Similarity of meaning* are discussed, the six experiments can be explained. Each dataset is conducted with the three kinds of instructions for *Similarity of meaning.*

Miller & Charles
Participants were presented with a localisation of the 30 word pairs in the experiment by Miller & Charles. Participants were asked to assign each word pair with a value on a five-point scale from 0 to 4. This experiment was conducted with each of the three kinds of instructions concerning *Similarity of meaning.* The official Dutch instructions can be found in appendices F, G, and H.

Rubenstein & Goodenough
Participants were presented with a localisation of the 65 word pairs in the experiment by Rubenstein & Goodenough. The original slips of paper were made containing these localisations. All 65 word pairs were placed on a different 85 by 55 mm slip of paper. Participants were then asked to order the slips of paper on similarity, assigning a value from 0 to 4 on a five-point scale. The official Dutch instructions can be found in appendices I, J, and K.

This concludes the subsection about the design of the experiments. For convenience, we will use abbreviations to refer to the six experiments. The abbreviation *Mc* will be used for the localisation of the dataset by Miller & Charles. *Rg* will be used for the localisation 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 the six experiments. For example, *McSim* means that the localisation of the dataset by Miller & Charles is meant with the instruction Similarity.

7.4 Participants

Pupils and teachers from five Dutch high schools participated. Pupils from the high school *Jacob-Roelandslyceum* in Boxtel participated in the pilot, and pupils from *'t Atrium* in Amersfoort, *Maurick College* in Vught, *RSG Trompmeesters* in Steenwijkerland and *'t Hooghe Landt* in Amersfoort participated in the experiments. 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 are: 40 for *McNo*, 40 for *McRel*, 52 for *McSim*, 26 for *RgNo*, 42 for *RgSim*, and 40 for *RgRel*.

7.5 Procedure participants

Each participant was presented with only one of the six experiments as described in subsection 7.3. Experiments were conducted in a school. An entire class did the same experiment at the same time. An effort was made to keep the group as silently as possible.

7.6 Procedure similarity measures

Each word pair from the localisation of the datasets by Miller & Charles and by Rubenstein & Goodenough was rated by the similarity measures *path*, *lch*, *wup*, *res*, *jcn*, and *lin*. In order to calculate these scores, Piek Vossen designed a package called *wordnet-tools*, [8] which makes it to possible to use Cornetto to calculate semantic similarity scores. There is one difference with the WordNet::Similarity package, which was designed by Pedersen, in how the similarity between words is calculated. Table 5 illustrates the issues:

Table 5: Synsets of *bos* 'forest' and *kerkhof* 'cemetery' are shown. The first column presents the synsets of *bos* 'forest', followed by the synsets of *kerkhof* 'cemetery' in the second column.

Synsets *bos*	**Synsets *kerkhof***
bos.n.01	kerkhof.n.01
bos.n.02	kerkhof.n.02

One way to calculate the similarity between two words could be to calculate all scores between all synsets of the two words of table 5, and then take the highest, which is what Pedersen implemented. However, we have opted to calculate all paths between the synsets of the two words. Out of these paths, the shortest path is chosen, and the score associated with this path will be representative for the similarity of the two words. We are aware that this choice might lead to different results than obtained by Pedersen. However, we claim that these differences will be small, due to the fact that the highest score often coincides with the shortest path. In addition, the default settings of the package were used. The maximum depth of the lexical semantic database, which is important for *lch*, was set to 31, because this is the maximum depth for nouns in Cornetto. Using this package, all similarity measures rated the two gold standards.

[8] The version from the 13th of July 2013 was used. After finishing this thesis, we found bugs in these tools. These bugs mainly effect the measures that are based on Information Content, which questions the reliability of the scores from these measures.

7.7 Analysis

In order to assess the correlation between datasets, at least two correlation measures are used in the literature. These are the Pearson product-moment correlation measure and the Spearman rank correlation ρ Spearman (1904). Miller and Charles (1991) uses the Pearson product-moment correlation measure to compare their own dataset to that of Rubenstein and Goodenough (1965). Patwardhan and Pedersen (2006) and Pedersen (2010) use the Spearman rank correlation measure to compare the results from human judgement against the results from similarity measures. However, one condition to conduct the Pearson product-moment correlation measure is that the data are normally distributed. This is not the case for both datasets. Results for the Kolmogorov-Smirnov test for normality indicated that both distributions did deviate significantly from a normal distribution, with $p < 0.01$ for both gold standards. This requirement does not need to be met to conduct the Spearman rank correlation measure. Hence, we will use the Spearman rank correlation measure to compare the results from the six experiments against the results from the similarity measures.

8 Results

The overall results can be found in table 6.

Table 6: The Spearman ρ is calculated by comparing all semantic similarity measures to the human judgements from all six experiments. The first columns shows the similarity measures, followed by the minimum and maximum of the experiments *McNo*, *McRel* and *McSim* in columns two and three. In the columns four and five, the same scores are shown for the experiments *RgNo*, *RgRel* and *RgSim*.

	Spearman ρ		**Spearman ρ**	
Measure	**min Mc**	**max Mc**	**min Rg**	**max Rg**
Path	0.741	0.788	0.654	0.712
Lch	0.741	0.788	0.654	0.712
Wup	0.365	0.373	0.175	0.237
Jcn	0.115	0.217	-0.197	-0.168
Lin	-0.179	-0.056	-0.253	-0.220
Res	-0.227	-0.093	-0.198	-0.134

In general, the results in table 6 show that the similarity measures *path* and *lch* seem to perform best for Dutch. In addition, measures based on Information Content do not seem to work well. This is a major difference with the English results, in which very little distinguished all six measures.

Table 7 presents the results per experiment.

Table 7: The Spearman ρ is shown by comparing all six similarity measure to all six experiments.

SM	McNo	McRel	McSim	RgNo	RgRel	RgSim
Path	0.777	0.741	0.788	0.712	0.654	0.707
Lch	0.777	0.741	0.788	0.712	0.654	0.707
Wup	0.365	0.373	0.369	0.237	0.175	0.233
Jcn	0.119	0.217	0.115	-0.197	-0.168	-0.171
Res	-0.209	-0.093	-0.227	-0.189	-0.134	-0.198
Lin	-0.179	-0.056	-0.178	-0.250	-0.220	-0.253

What stands out from table 7 is that the instructions did not have a big influence on the scores. The similarity measures seem to correlate best with the experiments in which the instruction was *Similarity*.

9 Discussion

In this section, the results will be discussed. We will start with a discussion on the measures that performed well, followed by an in-depth analysis of the measures that did not correlate well with human judgement. This is followed by a discussion on the effect of the different instructions. Finally, we will draw a general conclusion about the results.

The results from tables 6 and 7 show that the similarity measures *path* and *lch* seem to perform well. This is in accordance with the results from the previous English experiments.

More surprisingly are the low correlations of the measures *wup*, *res*, *jcn* and *lin*. However, all these measures require that the LCS is informative in order to be succesful. Closer inspection of the results of these measures showed that many of the least common subsumers of the word pairs were the root node. For example, for the similarity measure *res*, 55% of the least common subsumbers of the word pairs of the localisation of the word pairs by Rubenstein & Goodenough were the root node. To illustrate this, we present a comparison of the calculation of the scores of the similarity measure *res* for an English word pair in WordNet and its Dutch localisation in Cornetto. The English word pair that we will be using is *shore* and *woodland*, of which the Dutch localisation is *oever* and *woud*, respectively. The highest score for the English word pair is between the first senses of both words, whereas the shortest path for the Dutch word pair is also between the first senses of both words. The LCS for the Dutch word pair is the root node at depth 1, which is the second sense of *iets* ‘something’. However, for the English word pair, the LCS is ‘physical entity’ at depth 3 in WordNet. Because this occurs many times for the Dutch word pairs, the scores for these word pairs for the similarity measure *res* will all be exactly the same, because the score for this similarity measure is the Information Content of the LCS. This results in a very low correlation with human judgement. However for English, the scores are not the same, because the LCS for most word pairs are different.

Hence, the correlation with human judgement is better. This difference in the structure of the lexical semantic databases explains the poor performance of these measures for Dutch. This is likely caused by the fact that the top levels of Cornetto are quite general, whereas those of WordNet are more differentiated.

In addition, the limited difference between the different instructions is surprising. Despite instructions towards *Similarity*, *Relatedness* or even *No explanation*, very little distinguished the scores. We suggest the following interpretation of these results. Firstly, it seems like people have a general idea of similarity, which is very hard to change by providing specific instructions. Secondly, one can dount the reliability of the intuitions to quantify the similarity of two words. The intuitions seem to be more reliable when they are asked to say which word pair is more similar than another word pair, instead of asking to rate a single word pair.

To conclude, it is very hard to state that one similarity measure is better than another. Not only the formula determines the performance of a similarity measure, the structure of the semantic database also plays a crucial role. In order for a semantic similarity measure to work well, the elements that it requires need to be informative in the lexical semantic database. If a semantic measure depends on an is-a hierarchy, then this hierarchy must be informative in the lexical semantic database for the semantic measure to be succesful. It seems that WordNet is is much richer than Cornetto in these respects, in particular the diversity of the top levels, which is why most measures work well. However, a lexical semantic database in a different language, for example Cornetto, seems to have a different design of its top levels, which is why certain measures that require this do not work well.

10 Conclusion

The main focus of this thesis was upon getting a better understanding of the workings of semantic similarity measures. In order to do this, we did not only inspect the results from these measures in English using WordNet, but we also looked at the results using a different lexical semantic database in a different language, which is Cornetto for Dutch.

The research consisted of four steps. Firstly, we inspected the previous English experiments. Furthermore, we tried to replicate them to be sure that we fully understood the process. In addition, we created a Dutch gold standard. Finally, we inspected the correlations between the output from the semantic similarity measures using the Dutch lexical semantic database Cornetto and the newly created Dutch gold standard.

For English, we showed that a group of semantic similarity measures approached the human judgement in a similar way. Moreover, we stressed the importance of addressing every detail of the process that leads to the results by showing that even if the main properties are kept stable, variations in minor properties can lead to completely different outcomes. Furthermore, we presented our gold standard for Dutch and how it was created. In addition, we showed that not only the properties of a semantic similarity measure determine its performance, but that the structure of the lexical semantic database also plays a crucial role. More specifically, the measures *wup*, *res*, *jcn* and *lin* did not perform as well as in the English results, because the top levels of Cornetto are not as differentiated as in WordNet.

Acknowledgements

Many thanks go to my supervisors Jan Odijk and Piek Vossen. Moreover, I would like to thank the members of the Computational Lexicology and Terminology Web (CLTL) @ VU University Amsterdam for their feedback. In addition, thanks go to the co-authors of the paper ‘Offspring from Reproduction Problems: What Replication Failure Teaches Us’, which are Antske Fokkens, Marieke van Erp, Piek Vossen, Ted Pedersen and Nuno Freire. Many thanks also go to Rubén Izquierdo Beviá for his technical assistance. Finally, I 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

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Cramer, I. (2008). How well do semantic relatedness measures perform?: a meta-study. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1, pages 59–70.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

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

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

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., editor, *WordNet: An electronic lexical database*, pages 305–332. MIT Press.

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

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

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

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

Martin, W. and Maks, I. (2005). *Referentiebestand Nederlands documentatie.* INL.

Meng, L., Huang, R., and Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1):1–12.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

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

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308.

Oostdijk, N., Reynaert, M., Monachesi, P., Van Noord, G., Ordelman, R., Schuurman, I., and Vandeghinste, V. (2008). From D-Coi to SoNaR: a reference corpus for Dutch. In *LREC*.

Patwardhan, S. and Pedersen, T. (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*, pages 1–8, Trento, Italy.

Pedersen, T. (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)*, pages 329–332, Los Angeles, USA.

Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

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

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17–30.

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

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its applications to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

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

Spearman, C. (1904). Proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Strube, M. and Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Vossen, P. (1998). Introduction to EuroWordnet. *Computers and the Humanities*, 32(2-3):73–89.

Vossen, P. (2006). Cornetto: Een lexicaal-semantische database voor taaltechnologie. In *Dixit, special issue.*

Vossen, P., Hofman, K., de Rijke, M., Tjong Kim Sang, E., and Deschacht, K. (2007). The Cornetto Database: Architecture and User-Scenarios. In *Proceedings of 7th Dutch-Belgian Information Retrieval Workshop DIR2007*, Leuven, Belgium.

Vossen, P., Maks, I., Segers, R., and Van der Vliet, H. (2008). Integrating lexical units, synsets, and ontology in the Cornetto database. In *Proceedings on the Sixth International Conference on Language Resources and Evaluation (LREC 2008).*

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

A Dataset of 65 word pairs which was created by Rubenstein & Goodenough

Table 8: The 65 word pairs used by Rubenstein and Goodenough (1965). The first and the second column show the word pair, followed by the mean rating.

word	**word**	**value**
gem	jewel	3.94
midday	noon	3.94
automobile	car	3.92
cemetery	graveyard	3.88
cushion	pillow	3.84
boy	lad	3.82
cock	rooster	3.68
implement	tool	3.66
forest	woodland	3.65
coast	shore	3.60
autograph	signature	3.59
journey	voyage	3.58
serf	slave	3.46
grin	smile	3.46
glass	tumbler	3.45
cord	string	3.41
hill	mound	3.29
magician	wizard	3.21
furnace	stove	3.11
asylum	madhouse	3.04
brother	monk	2.74
food	fruit	2.69
bird	cock	2.63
bird	crane	2.63
oracle	sage	2.61
sage	wizard	2.46
brother	lad	2.41
crane	implement	2.37
magician	oracle	1.82
glass	jewel	1.78
cemetery	mound	1.69
car	journey	1.55
hill	woodland	1.48
crane	rooster	1.41

Continued on next page

Table 8 – *Continued from previous page*

word	**word**	**value**
furnace	implement	1.37
coast	hill	1.26
bird	woodland	1.24
shore	voyage	1.22
cemetery	woodland	1.18
food	rooster	1.09
forest	graveyard	1.00
lad	wizard	0.99
mound	shore	0.97
automobile	cushion	0.97
boy	sage	0.96
monk	oracle	0.91
shore	woodland	0.90
grin	lad	0.88
coast	forest	0.85
asylum	cemetery	0.79
monk	slave	0.57
cushion	jewel	0.45
boy	rooster	0.44
glass	magician	0.44
graveyard	madhouse	0.42
asylum	monk	0.39
asylum	fruit	0.19
grin	implement	0.18
mound	stove	0.14
automobile	wizard	0.11
autograph	shore	0.06
fruit	furnace	0.05
noon	string	0.04
rooster	voyage	0.04
cord	smile	0.00

B Dataset of 30 word pairs which was created by Miller & Charles

Table 9: The 30 word pairs used by Miller and Charles (1991). This set is a subset from the set from Rubenstein and Goodenough (1965). For each line, the left and middle column present the word pair, followed by the mean rating.

word	**word**	**value**
car	automobile	3.92
gem	jewel	3.84
journey	voyage	3.84
boy	lad	3.76
coast	shore	3.70
asylum	madhouse	3.61
magician	wizard	3.50
midday	noon	3.42
furnace	stove	3.11
food	fruit	3.08
bird	cock	3.05
bird	crane	2.97
tool	implement	2.95
brother	monk	2.82
lad	brother	1.66
crane	implement	1.68
journey	car	1.16
monk	oracle	1.10
cemetery	woodland	0.95
food	rooster	0.89
coast	hill	0.87
forest	graveyard	0.84
shore	woodland	0.63
monk	slave	0.55
coast	forest	0.42
lad	wizard	0.42
chord	smile	0.13
glass	magician	0.11
rooster	voyage	0.08
noon	string	0.00

C Word sense disambiguation for 49 nouns from datasets Miller & Charles and Rubenstein & Goodenough

Table 10: The word sense disambiguation for the set of 49 nouns used in the word pairs of the studies by Miller and Charles (1991) and Rubenstein and Goodenough (1965) are shown. In the first column, the entry in Wordnet is shown, followed by the sense number in the second column. The third column presents an explanation if not the first sense listed in WordNet is chosen.

Wordnet entry	sense number	explanation
asylum	2	Due to the high correlation scores from human judgement between 'asylum' and 'madhouse' in both studies, we assume that the meaning of 'a hospital for mentally incompetent or unbalanced person' is meant.
autograph	2	Due to the high correlation in human judgement in the study by Rubenstein & Goodenough (1965) with 'signature', we believe that the second meaning of 'a person's own signature' is meant here.
automobile	1	none
bird	1	none
boy	1	none
brother	1	Due to the high correlation between 'brother' and 'monk' of 2.74, we believe that the religious meaning is meant.
car	1	none
cemetery	1	none
cock	4	Due to the high correlation between 'bird' and 'cock' in both studies, we believe that the meaning of 'adult male chicken' is meant here.
cord	1	The correlation between 'chord' and 'smile' is very low in the study by Rubenstein & Goodenough (1965), which is why it is very difficult to decide which meaning is meant. WordNet lists two meanings: (1) 'a straight line connecting two points on a curve', and meaning (2) 'a combination of three or more notes that blend harmoniously when sounded together'. We believe that the second meaning is the most common, which is why we opted for this meaning.
chord	2	This meaning is very uncommon, which is why we have chosen for the second meaning.
coast	1	none

Continued on next page

Table 10 – *Continued from previous page*

Wordnet entry	**sense number**	**explanation**
crane	5	Due to the correlation of 2.63 between ‘bird’ and ‘crane’ and the correlation of 2.37 between ‘crane’ and ‘implement’ , we believe that the meaning of ‘large long-necked wading bird of marshes and plains in many parts of the world’ AND ‘lifts and moves heavy objects lifting tackle is suspended from a pivoted boom that rotates around a vertical axis’ should be present in the Dutch localisation. Finally, ‘ezel’ (donkey) was chosen, because this has a meaning which is related to both ‘implement’ and ‘bird’.
cushion	3	Due to the high correlation with ‘pillow’ in the study by Rubenstein & Goodenough (1965), we believe that the meaning of ‘a soft bag filled with air or a mass of padding such as feathers or foam rubber etc.’ is meant here.
food	1	none
forest	2	Due to the high human correlation with ‘woodland’ in the study by Rubenstein & Goodenough, we have opted for the second sense.
fruit	1	none
furnace	1	none
gem	2	Due to the high correlation scores from human judgement between ‘gem’ and ‘jewel’ in both studies, we assume that the meaning of ‘a crystalline rock that can be cut and polished for jewellery’ is meant.
glass	1	none
graveyard	1	none
grin	1	none
hill	1	none
implement	1	none
jewel	1	none
journey	1	none
lad	1	none
madhouse	1	none
magician	2	Due to the high human correlation with ‘wizard’ in both studies, we have opted for the second sense.
midday	1	none
monk	1	none
mound	2	Due to the fact that the first meaning is culturally dependent (baseball domain), we have opted for the sense of ‘natural mound’.
noon	1	none
oracle	1	none
pillow	1	none
rooster	1	none
sage	1	none
serf	1	none

Continued on next page

Table 10 – *Continued from previous page*

Wordnet entry	**sense number**	**explanation**
shore	1	none
signature	1	none
slave	1	none
smile	1	none
stove	1	none
string	1	none
tool	2	The first sense is religious. Due to the high correlation with 'implement' in both studies, We believe the second sense is more appropriate.
tumbler	2	Due to the high correlation with 'glass' in the study by Rubenstein & Goodenough (1965), we believe the meaning of 'a glass with a flat bottom but no handle or stem originally had a round bottom' is intended.
voyage	2	We believe that the meaning of 'an act of travailing by water' is too specific, which is why we prefer the second more general sense.
wizard	2	Due to the high correlation between 'wizard' and 'magician' in both studies, we assume that meaning of 'one who practices magic or sorcery' is meant here.
woodland	1	none

D Dutch localisation

Table 11: The Dutch localisations for the set of 49 nouns used in the word pairs of the studies by Miller and Charles (1991) and Rubenstein and Goodenough (1965) are shown. In the first column, the entry in Cornetto is shown, followed by the sense number in the second column. The third column presents an explanation if needed.

English noun	Dutch localisation	explanation
asylum	inrichting	none
autograph	handtekening	none
automobile	wagen	none
bird	vogel	none
boy	jongen	none
brother	broeder	none
car	auto	none
cemetery	begraafplaats	none
cock	kip	Due to the fact that there are not two common Dutch synonyms for ‘cock’ and ‘rooster’, we opted to replace these two words by ’kip’ and ’hen’, which are two Dutch words for female chickens.
cord	koord	none
chord	akkoord	none
coast	kust	none
crane	ezel	none
cushion	bed	Due to the fact that there are not two Dutch synonyms for ‘cushion’ and ‘pillow’, we opted to use two words for ‘bed’ (bed), which are ’bed’ and ’nest’.
food	voedsel	none
forest	bos	none
fruit	fruit	none
furnace	oven	none
gem	edelsteen	none
glass	glas	none
graveyard	kerkhof	none
grin	grijns	none
hill	berg	none
implement	werktuig	none
jewel	juweel	none
journey	tocht	none
lad	knul	none
madhouse	gekkenhuis	none
magician	magier	none

Continued on next page

Table 11 – *Continued from previous page*

English noun	**Dutch localisation**	**explanation**
midday	ochtend	Due to the fact that there are not two Dutch synonyms for ‘midday’ and ‘noon’, we opted to localise these words by ’ochtend’ and ’morgen’, the two dutch words for ‘morning’.
monk	monnik	none
mound	heuvel	none
noon	morgen	Due to the fact that there are not two Dutch synonyms for ‘midday’ and ‘noon’, we opted to localise these words by ’ochtend’ and ’morgen’, the two dutch words for ‘morning’.
oracle	orakel	none
pillow	nest	Due to the fact that there are not two Dutch synonyms for ‘cushion’ and ‘pillow’, we opted to use two words for ‘bed’ (bed), which are ’bed’ and ’nest’.
rooster	hen	Due to the fact that there are not two common Dutch synonyms for ‘cock’ and ‘rooster’, we opted to replace these two words by ’kip’ and ’hen’, which are two Dutch words for female chickens.
sage	wijsgeer	none
serf	lijfeigene	none
shore	oever	none
signature	signatuur	none
slave	slaaf	none
smile	glimlach	none
stove	fornuis	none
string	draadje	none
tool	gereedschap	none
tumbler	drinkbeker	none
voyage	reis	none
wizard	tovenaar	none
woodland	woud	none

E Comparison relative frequency between English nouns and their Dutch localisations

Table 12: The comparison of the relative frequency, or RF, of the English nouns and their Dutch localisations are shown. In the first three columns, the English nouns with their relative frequency in SemCor and their relative frequency class are shown, respectively. A relative frequency of 0.0182% means that the word occupies 0,0182% of the corpus. In column four until six, the Dutch translocations are shown with their relative frequencies in SoNar and their relative frequency class, respectively. 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 is set to True.

Noun(EN)	RF	RF class	Noun(DU)	RF	RF class	Same RF class?
asylum	0.0	low	inrichting	0.01404	low	True
autograph	0.0	low	handtekening	0.01035	low	True
automobile	0.01607	middle	wagen	0.03154	middle	True
bird	0.03107	middle	vogel	0.02237	middle	True
boy	0.15428	high	jongen	0.05216	high	True
brother	0.03857	middle	broeder	0.00456	low	False
car	0.07607	high	auto	0.094	high	True
cemetery	0.00429	low	begraafplaats	0.00333	low	True
cock	0.0	low	kip	0.01544	middle	False
cord	0.00536	low	koord	0.00165	low	True
chord	0.00214	low	akkoord	0.04085	middle	False
coast	0.02143	middle	kust	0.02116	middle	True
crane	0.0	low	ezel	0.00402	low	True
cushion	0.0	low	bed	0.03022	middle	False
food	0.03107	middle	voedsel	0.0201	middle	True
forest	0.00214	low	bos	0.02669	middle	False
fruit	0.01071	low	fruit	0.01221	low	True
furnace	0.00107	low	oven	0.00707	low	True
gem	0.0	low	edelsteen	0.00109	low	True
glass	0.02357	middle	glas	0.02224	middle	True
graveyard	0.00321	low	kerkhof	0.00313	low	True
grin	0.00643	low	grijns	0.00079	low	True
hill	0.03535	middle	berg	0.01856	middle	True
implement	0.00214	low	werktuig	0.00171	low	True

Continued on next page

Table 12 – *Continued from previous page*

Noun(EN)	**RF**	**RF class**	**Noun(DU)**	**RF**	**RF class**	**Same RF class?**
jewel	0.00214	low	juweel	0.00646	low	True
journey	0.01821	middle	tocht	0.01088	low	False
lad	0.00429	low	knul	0.00041	low	True
madhouse	0.0	low	gekkenhuis	0.00046	low	True
magician	0.0	low	magier	0.0	low	True
midday	0.00107	low	ochtend	0.01263	low	True
monk	0.00536	low	monnik	0.00447	low	True
mound	0.00107	low	heuvel	0.00734	low	True
noon	0.00857	low	morgen	0.00417	low	True
oracle	0.0	low	orakel	0.00057	low	True
pillow	0.00857	low	nest	0.00748	low	True
rooster	0.00321	low	hen	0.00033	low	True
sage	0.00107	low	wijsgeer	0.00035	low	True
serf	0.0	low	lijfeigene	0.0002	low	True
shore	0.01928	middle	oever	0.00514	low	False
signature	0.00107	low	signatuur	0.00148	low	True
slave	0.01928	middle	slaaf	0.00643	low	False
smile	0.03107	middle	glimlach	0.00464	low	False
stove	0.00536	low	fornuis	0.00129	low	True
string	0.00536	low	draadje	0.00131	low	True
tool	0.00536	low	gereedschap	0.00313	low	True
tumbler	0.00107	low	drinkbeker	0.00012	low	True
voyage	0.00214	low	reis	0.03439	middle	False
wizard	0.0	low	tovenaar	0.00196	low	True
woodland	0.0	low	woud	0.00448	low	True

F Experiment Miller & Charles with instruction Similarity

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is om voor 30 woordparen aan te geven hoe gelijk jij de betekenis van twee woorden vindt.

Hoe gaat dit in z'n werk?

(2) Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. Dit vul je in in de derde kolom genaamd **waarde**. In dit geval hebben wij dit woordpaar een waarde van 4 toegekend.

Eerste woord	Tweede woord	Waarde
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

(4) Twee woorden hebben een hoge "gelijkheid van betekenis" als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen. Als je iets een **fiets** noemt dan kun je het ook een **rijwiel** noemen. Om je te helpen bij het bepalen van de waardes, geven wij een aantal voorbeelden.

Waarde 4 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen en ook andersom:

Voorbeeld waarde 4: **fiets** en **rijwiel**

uitleg: '**fiets**' kan je gebruiken in de plaats van '**rijwiel**' en andersom.

Waarde 3 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom:

waarde 3: **aardappelmesje** en **mes**

uitleg '**mes**' kan je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

Waarde 2 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom en het is wat algemener.

waarde 2: **aardappelmesje** en **eetgerei**

uitleg: '**eetgerei**' kan je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

waarde 2: **vliegtuig** en **auto**

uitleg: '**vliegtuig**' en '**auto**' kan je beide gebruiken in plaats van '**vervoersmiddel**'.

Waarde 1 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom en het is nog algemener.

waarde 1: **aardappelmesje** en **materiaal**

uitleg: '**materiaal**' kun je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

waarde 1: **mens** en **kat**

uitleg: '**mens**' en '**kat**' kun je beide gebruiken in plaats van '**levende wezens**'.

Waarde 0 vinden we als je één van de twee woorden niet kan gebruiken in plaats van het andere woord of als het heel algemeen is.

waarde 0: **aardappelmesje** en **iets**

uitleg: '**iets**' kan je gebruiken in plaats van '**aardappelmesje**', maar het is te algemeen.

waarde 0: **komkommer** en **professor**

uitleg: de relatie tussen de woorden is te algemeen.

(5) Dan ben jij nu aan de beurt. Geef alsjeblieft op de volgende pagina aan elk van de woordparen een waarde. De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

Alvast bedankt voor je medewerking.

Eerste Woord	Tweede Woord	Waarde
ezel	werktuig	
voedsel	hen	
monnik	slaaf	
oven	fornuis	
morgen	draadje	
tocht	reis	
voedsel	fruit	
jongen	knul	
hen	reis	
begraafplaats	woud	
kust	berg	
knul	tovenaar	
edelsteen	juweel	
inrichting	gekkenhuis	
monnik	orakel	
magier	tovenaar	
kust	bos	
kust	oever	
tocht	auto	
oever	woud	
broeder	monnik	
bos	kerkhof	
auto	wagen	
glas	magier	
akkoord	glimlach	
ochtend	morgen	
gereedschap	werktuig	
vogel	kip	
knul	broeder	
vogel	ezel	

G Experiment Miller & Charles with instruction Relatedness

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is om voor 30 woordparen aan te geven hoe gelijk jij de betekenis van twee woorden vindt.

Hoe gaat dit in z'n werk?

(2) Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. Dit vul je in in de derde kolom genaamd **waarde**. In dit geval hebben wij dit woordpaar een waarde van 4 toegekend.

Eerste woord	Tweede woord	Waarde
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

(4) Twee woorden hebben een hoge "gelijkheid van betekenis" als je ze beide in dezelfde situatie kunt tegenkomen. **Fietspad** en **fiets** hebben bijvoorbeeld een hoge "gelijkheid van betekenis", omdat je verwacht dat je fietsen op een fietspad tegenkomt.

Om je te helpen bij het bepalen van de waardes, geven wij een aantal voorbeelden.

waarde 0: **komkommer** en **professor**

uitleg: deze woorden hebben een lage "gelijkheid van betekenis", omdat het zeer onwaarschijnlijk is dat "**komkommer**" en "**professor**" in dezelfde situatie voorkomen.

waarde 1: **probleem** en **vliegveld**

uitleg: het is iets waarschijnlijker dat "**probleem**" en "**vliegveld**" samen voorkomen, omdat er wel eens problemen op vliegvelden zijn.

waarde 2: **auto** en **vlucht**

uitleg: **Auto** en **vlucht** hebben al een wat hogere "gelijkheid van betekenis" omdat je in een auto kunt vluchten.

waarde 3:**computer** en **internet**

uitleg: het is zeer waarschijnlijk dat "**computer**" en "**internet**" samen voorkomen, waardoor de "gelijkheid van betekenis" hoog is.

Waarde 4: **fiets** en **rijwiel**

uitleg: deze woorden zijn synoniemen, waardoor de hoogste "gelijkheid van betekenis" wordt toegekend.

(5) Dan ben jij nu aan de beurt. Geef alsjeblieft op de volgende pagina aan elk van de woordparen een waarde. De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

Alvast bedankt voor je medewerking.

Eerste Woord	**Tweede Woord**	**Waarde**
ezel	werktuig	
voedsel	hen	
monnik	slaaf	
oven	fornuis	
morgen	draadje	
tocht	reis	
voedsel	fruit	
jongen	knul	
hen	reis	
begraafplaats	woud	
kust	berg	
knul	tovenaar	
edelsteen	juweel	
inrichting	gekkenhuis	
monnik	orakel	
magier	tovenaar	
kust	bos	
kust	oever	
tocht	auto	
oever	woud	
broeder	monnik	
bos	kerkhof	
auto	wagen	
glas	magier	
akkoord	glimlach	
ochtend	morgen	
gereedschap	werktuig	
vogel	kip	
knul	broeder	
vogel	ezel	

H Experiment Miller & Charles with no instruction

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is om 30 woordparen een waarde te geven op basis van “gelijkheid van betekenis”.

Hoe gaat dit in z’n werk?

(2) Bij dit experiment krijg je vrijwel geen uitleg over wat we bedoelen met “gelijkheid van betekenis”. De enige tip die we geven is dat twee woorden de hoogste “gelijkheid van betekenis” hebben wanneer deze synoniemen zijn van elkaar. Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. Dit vul je in in de derde kolom genaamd **waarde**. In dit geval hebben wij dit woordpaar een waarde van 4 toegekend.

Eerste woord	Tweede woord	Waarde
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de woorden een lage “gelijkheid van betekenis” hebben en 4 dat de woorden een hoge “gelijkheid van betekenis” hebben.

(4) Dan ben jij nu aan de beurt. Geef alsjeblieft op de volgende pagina aan elk van de woordparen een waarde.

Alvast bedankt voor je medewerking.

Eerste Woord	Tweede Woord	Waarde
ezel	werktuig	
voedsel	hen	
monnik	slaaf	
oven	fornuis	
morgen	draadje	
tocht	reis	
voedsel	fruit	
jongen	knul	
hen	reis	
begraafplaats	woud	
kust	berg	
knul	tovenaar	
edelsteen	juweel	
inrichting	gekkenhuis	
monnik	orakel	
magier	tovenaar	
kust	bos	
kust	oever	
tocht	auto	
oever	woud	
broeder	monnik	
bos	kerkhof	
auto	wagen	
glas	magier	
akkoord	glimlach	
ochtend	morgen	
gereedschap	werktuig	
vogel	kip	
knul	broeder	
vogel	ezel	

I Experiment Rubenstein & Goodenough with instruction Similarity

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is dat je een stapel van 65 kaartjes op volgorde gaat leggen van 'gelijkheid van betekenis'. We zullen zo uitleggen wat we daarmee bedoelen. Wanneer je de kaartjes op volgorde hebt gelegd, vragen we je om aan elk kaartje een waarde te geven.

Wat is 'gelijkheid van betekenis'?

(2) Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. In dit geval hebben wij dit woordpaar de maximale waarde van 4 toegekend.

Eerste woord	Tweede woord	Waarde
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

(4) Twee woorden hebben een hoge "gelijkheid van betekenis" als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen. Als je iets een **fiets** noemt dan kun je het ook een **rijwiel** noemen. Om je te helpen bij het bepalen van de waardes, geven wij een aantal voorbeelden.

Waarde 4 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen en ook andersom:

Voorbeeld waarde 4: **fiets** en **rijwiel**

uitleg: '**fiets**' kan je gebruiken in de plaats van '**rijwiel**' en andersom.

Waarde 3 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom:

waarde 3: **aardappelmesje** en **mes**

uitleg '**mes**' kan je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

Waarde 2 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom en het is wat algemener.

waarde 2: **aardappelmesje** en **eetgerei**

uitleg: '**eetgerei**' kan je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

waarde 2: **vliegtuig** en **auto**

uitleg: '**vliegtuig**' en '**auto**' kan je beide gebruiken in plaats van '**vervoersmiddel**'.

Waarde 1 vinden we als je één van de twee woorden in plaats van het andere woord kunt gebruiken om iets te noemen, maar niet andersom en het is nog algemener.

waarde 1: **aardappelmesje** en **materiaal**

uitleg: '**materiaal**' kun je gebruiken in plaats van '**aardappelmesje**', maar niet andersom.

waarde 1: **mens** en **kat**

uitleg: '**mens**' en '**kat**' kun je beide gebruiken in plaats van '**levende wezens**'.

Waarde 0 vinden we als je één van de twee woorden niet kan gebruiken in plaats van het andere woord of als het heel algemeen is.

waarde 0: **aardappelmesje** en **iets**

uitleg: '**iets**' kan je gebruiken in plaats van '**aardappelmesje**', maar het is te algemeen.

waarde 0: **komkommer** en **professor**

uitleg: de relatie tussen de woorden is te algemeen.

(5) Dan ben jij nu aan de beurt. Leg alsjeblieft je stapel met kaartjes op volgorde van gelijkheid van betekenis. Wanneer je hiermee klaar bent, vragen we je om te bepalen welk deel van je kaartjes je de waarde 4 zou geven, welke de waarde 3, welke de waarde 2, welke de waarde 1 en welke de waarde 0. Elk kaartje heeft naast de twee woorden ook een uniek **identificatienummer**. We zouden je willen vragen om op het volgende blad aan te geven welke kaartjes je welke waarde hebt gegeven. De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

TIP: als je een woord niet kent, leg het kaartje dan weg.

TIP: het is het handigste om de kolommen te maken van kaartjes die je waarde 0 geeft, waarde 1 etc. op je tafel.

Alvast bedankt voor je medewerking.

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 4 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 3 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 2 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 1 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 0 hebt gegeven:

J Experiment Rubenstein & Goodenough with instruction Relatedness

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is dat je een stapel van 65 kaartjes op volgorde gaat leggen van 'gelijkheid van betekenis'. We zullen zo uitleggen wat we daarmee bedoelen. Wanneer je de kaartjes op volgorde hebt gelegd, vragen we je om aan elk kaartje een waarde te geven.

Wat is 'gelijkheid van betekenis'?

(2) Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. In dit geval hebben wij dit woordpaar de maximale waarde van 4 toegekend.

Eerste woord	Tweede woord	Waarde
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

(4) Twee woorden hebben een hoge "gelijkheid van betekenis" als je ze beide in dezelfde situatie kunt tegenkomen. **Fietspad** en **fiets** hebben bijvoorbeeld een hoge "gelijkheid van betekenis", omdat je verwacht dat je fietsen op een fietspad tegenkomt.

Om je te helpen bij het bepalen van de waardes, geven wij een aantal voorbeelden.

waarde 0: **komkommer** en **professor**

uitleg: deze woorden hebben een lage "gelijkheid van betekenis", omdat het zeer onwaarschijnlijk is dat "**komkommer**" en "**professor**" in dezelfde situatie voorkomen.

waarde 1: **probleem** en **vliegveld**

uitleg: het is iets waarschijnlijker dat "**probleem**" en "**vliegveld**" samen voorkomen, omdat er wel eens problemen op vliegvelden zijn.

waarde 2: **auto** en **vlucht**

uitleg: **Auto** en **vlucht** hebben al een wat hogere "gelijkheid van betekenis" omdat je in een auto kunt vluchten.

waarde 3:**computer** en **internet**

uitleg: het is zeer waarschijnlijk dat "**computer**" en "**internet**" samen voorkomen, waardoor de "gelijkheid van betekenis" hoog is.

Waarde 4: **fiets** en **rijwiel**

uitleg: deze woorden zijn synoniemen, waardoor de hoogste "gelijkheid van betekenis" wordt toegekend.

(5) Dan ben jij nu aan de beurt. Leg alsjeblieft je stapel met kaartjes op volgorde van gelijkheid van betekenis. Wanneer je hiermee klaar bent, vragen we je om te bepalen welk deel van je kaartjes je de waarde 4 zou geven, welke de waarde 3, welke de waarde 2, welke de waarde 1 en welke de waarde 0. Elk kaartje heeft naast de twee woorden ook een uniek **identificatienummer**. We zouden je willen vragen om op het volgende blad aan te geven welke kaartjes je welke waarde hebt gegeven. De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de betekenis van de woorden helemaal niet gelijk is en 4 dat de betekenis van de woorden zo goed als gelijk is.

TIP: als je een woord niet kent, leg het kaartje dan weg.

TIP: het is het handigste om de kolommen te maken van kaartjes die je waarde 0 geeft, waarde 1 etc. op je tafel.

Alvast bedankt voor je medewerking.

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 4 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 3 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 2 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 1 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 0 hebt gegeven:

K Experiment Rubenstein & Goodenough with no instruction

The contents of this appendix can be found on the next page.

Hallo, welkom bij dit experiment, alvast bedankt dat je mee wilt doen!

Voordat je begint, zouden we je willen vragen om wat gegevens over jezelf in te vullen.

Leerlingnummer:

(1) Het doel van dit experiment is dat je een stapel van 65 kaartjes op volgorde gaat leggen van 'gelijkheid van betekenis'. Wanneer je de kaartjes op volgorde hebt gelegd, vragen we je om aan elk kaartje een waarde te geven.

(2) Bij dit experiment krijg je vrijwel geen uitleg over wat we bedoelen met "gelijkheid van betekenis". De enige tip die we geven is dat twee woorden de hoogste "gelijkheid van betekenis" hebben wanneer deze synoniemen zijn van elkaar. Als voorbeeld, kun je kijken naar de tabel hieronder. Je ziet in de eerste kolom het woord **fiets** en in de tweede kolom het woord **rijwiel**. Het is dan jouw taak om te bepalen hoe gelijk de betekenis is van deze twee woorden. In dit geval hebben wij dit woordpaar de maximale waarde van 4 toegekend.

Eerste woord	**Tweede woord**	**Waarde**
fiets	rijwiel	4

(3) De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4, waarbij 0 betekent dat de "gelijkheid van betekenis" laag is en 4 dat de "gelijkheid van betekenis" hoog is.

(4) Dan ben jij nu aan de beurt. Leg alsjeblieft je stapel met kaartjes op volgorde van "gelijkheid van betekenis". Wanneer je hiermee klaar bent, vragen we je om te bepalen welk deel van je kaartjes je de waarde 4 zou geven, welke de waarde 3, welke de waarde 2, welke de waarde 1 en welke de waarde 0. Elk kaartje heeft naast de twee woorden ook een uniek **identificatienummer**. We zouden je willen vragen om op het volgende blad aan te geven welke kaartjes je welke waarde hebt gegeven. De waardes die je kan toekennen aan een woordpaar zijn 0 of 1 of 2 of 3 of 4.

Alvast bedankt voor je medewerking.

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 4 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 3 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 2 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 1 hebt gegeven:

Vul hier alsjeblieft de **identificatienummers** in van de kaartjes die je de waarde 0 hebt gegeven:

L Preprint paper entitled 'Offspring from Reproduction Problems: What Replication Failure Teaches Us'

The contents of this appendix can be found on the next page.

Offspring from Reproduction Problems: What Replication Failure Teaches Us

Antske Fokkens and **Marieke van Erp**
The Network Institute
VU University Amsterdam
Amsterdam, The Netherlands
{a.s.fokkens,m.g.j.van.erp}@vu.nl

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

Ted Pedersen
Dept. of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

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

Nuno Freire
The European Library
The Hague, The Netherlands
nfreire@gmail.com

Abstract

Repeating experiments is an important instrument in the scientific toolbox to validate previous work and build upon existing work. We present two concrete use cases involving key techniques in the NLP domain for which we show that reproducing results is still difficult. We show that the deviation that can be found in reproduction efforts leads to questions about how our results should be interpreted. Moreover, investigating these deviations provides new insights and a deeper understanding of the examined techniques. We identify five aspects that can influence the outcomes of experiments that are typically not addressed in research papers. Our use cases show that these aspects may change the answer to research questions leading us to conclude that more care should be taken in interpreting our results and more research involving systematic testing of methods is required in our field.

1 Introduction

Research is a collaborative effort to increase knowledge. While it includes validating previous approaches, our experience is that most research output in our field focuses on presenting new approaches, and to a somewhat lesser extent building upon existing work.

In this paper, we argue that the value of research that attempts to replicate previous approaches goes beyond simply *validating* what is already known. It is also an essential aspect for *building upon* existing approaches. Especially when validation fails or variations in results are found, systematic testing helps to obtain a clearer picture of both the approach itself and of the meaning of state-of-the-art results leading to a *better insight* into the quality of new approaches in relation to previous work.

We support our claims by presenting two use cases that aim to reproduce results of previous work in two key NLP technologies: measuring WordNet similarity and Named Entity Recognition (NER). Besides highlighting the difficulty of repeating other researchers' work, new insights about the approaches emerged that were not presented in the original papers. This last point shows that reproducing results is not merely part of good practice in science, but also an essential part in gaining a better understanding of the methods we use. Likewise, the problems we face in reproducing previous results are not merely frustrating inconveniences, but also pointers to research questions that deserve deeper investigation.

We investigated five aspects that cause experimental variation that are not typically described in publications: **preprocessing** (e.g. tokenisation), **experimental setup** (e.g. splitting data for cross-validation), **versioning** (e.g. which version of WordNet), **system output** (e.g. the exact features used for individual tokens in NER), and **system variation** (e.g. treatment of ties).

As such, reproduction provides a platform for systematically testing individual aspects of an approach that contribute to a given result. What is the influence of the size of the dataset, for example? How does using a different dataset affect the results? What is a reasonable divergence between different runs of the same experiment? Finding answers to these questions enables us to better interpret our state-of-the-art results.

Moreover, the experiments in this paper show that even while strictly trying to replicate a previous experiment, results may vary up to a point where they lead to different answers to the main question addressed by the experiment. The WordNet similarity experiment use case compares the performance of different similarity measures. We will show that the answer as to which measure works best changes depending on factors such as the gold standard used, the strategy towards part-of-speech or the ranking coefficient, all aspects that are typically not addressed in the literature.

The main contributions of this paper are the following:

1) An in-depth analysis of two reproduction use cases in NLP

2) New insights into the state-of-the-art results for WordNet similarities and NER, found because of problems in reproducing prior research

3) A categorisation of aspects influencing reproduction of experiments and suggestions on testing their influence systematically

The code, data and experimental setup for the WordNet experiments are available at `http://github.com/antske/WordNetSimilarity`, and for the NER experiments at `http://github.com/Mvanerp/NER`. The experiments presented in this paper have been repeated by colleagues not involved in the development of the software using the code included in these repositories. The remainder of this paper is structured as follows. In Section 2, previous work is discussed. Sections 3 and 4 describe our real-world use cases. In Section 5, we present our observations, followed by a more general discussion in Section 6. In Section 7, we present our conclusions.

2 Background

This section provides a brief overview of recent work addressing reproduction and benchmark results in computer science related studies and discusses how our research fits in the overall picture.

Most researchers agree that validating results entails that a method should lead to the same overall conclusions rather than producing the exact same numbers (Drummond, 2009; Dalle, 2012; Buchert and Nussbaum, 2012, etc.). In other words, we should strive to *reproduce* the same answer to a research question by different means, perhaps by re-implementing an algorithm or evaluating it on a new (in domain) data set. *Replication* has a somewhat more limited aim, and simply involves running the exact same system under the same conditions in order to get the exact same results as output.

According to Drummond (2009) replication is not interesting, since it does not lead to new insights. On this point we disagree with Drummond (2009) as replication allows us to: 1) validate prior research, 2) improve on prior research without having to rebuild software from scratch, and 3) compare results of reimplementations and obtain the necessary insights to perform reproduction experiments. The outcome of our use cases confirms the statement that deeper insights into an approach can be obtained when all resources are available, an observation also made by Ince et al. (2012).

Even if exact replication is not a goal many strive for, Ince et al. (2012) argue that insightful reproduction can be an (almost) impossible undertaking without the source code being available. Moreover, it is not always clear where replication stops and reproduction begins. Dalle (2012) distinguishes levels of reproducing results related to how close they are to the original work and how each contributes to research. In general, an increasing awareness of the importance of reproduction research and open code and data can be observed based on publications in high-profile journals (e.g. Nature (Ince et al., 2012)) and initiatives such as myExperiment.[1]

Howison and Herbsleb (2013) point out that, even though this is important, often not enough (academic) credit is gained from making resources available. What is worse, the same holds for research that investigates existing methods rather than introducing new ones, as illustrated by the question that is found on many review forms 'how novel is the presented approach?'. On the other hand, initiatives for journals addressing exactly this issue (Neylon et al., 2012) and tracks focusing on results verification at conferences such as VLDB[2] show that this opinion is not universal.

A handful of use cases on reproducing or replicating results have been published. Louridas and Gousios (2012) present a use case revealing that source code alone is not enough for reproducing

[1] `http://www.myexperiment.org`

[2] `http://www.vldb.org/2013/`

results, a point that is also made by Mende (2010) who provides an overview of all information required to replicate results.

The experiments in this paper provide use cases that confirm the points brought out in the literature mentioned above. This includes both observations that a detailed level of information is required for truly insightful reproduction research as well as the claim that such research leads to better understanding of our techniques. Furthermore, the work in this paper relates to Bikel (2004)'s work. He provides all information needed in addition to Collins (1999) to replicate Collins' benchmark results. Our work is similar in that we also aim to fill in the blanks needed to replicate results. It must be noted, however, that the use cases in this paper have a significantly smaller scale than Bikel's.

Our research distinguishes itself from previous work, because it links the challenges of reproduction to what they mean for reported results beyond validation. Ruml (2010) mentions variations in outcome as a reason not to emphasise comparisons to benchmarks. Vanschoren et al. (2012) propose to use experimental databases to systematically test variations for machine learning, but neither links the two issues together. Raeder et al. (2010) come closest to our work in a critical study on the evaluation of machine learning. They show that choices in the methodology, such as data sets, evaluation metrics and type of cross-validation can influence the conclusions of an experiment, as we also find in our second use case. However, they focus on the problem of evaluation and recommendations on how to achieve consistent reproducible results. Our contribution is to investigate how much results vary. We cannot control how fellow researchers carry out their evaluation, but if we have an idea of the variations that typically occur within a system, we can better compare approaches for which not all details are known.

3 WordNet Similarity Measures

Patwardhan and Pedersen (2006) and Pedersen (2010) present studies where the output of a variety of WordNet similarity and relatedness measures are compared. They rank Miller and Charles (1991)'s set (henceforth "*mc-set*") of 30 word pairs according to their semantic relatedness with several WordNet similarity measures.

Each measure ranks the *mc-set* of word pairs and these outputs are compared to Miller and Charles (1991)'s gold standard based on human rankings using the Spearman's Correlation Coefficient (Spearman, 1904, ρ). Pedersen (2010) also ranks the original set of 65 word pairs ranked by humans in an experiment by Rubenstein and Goodenough (1965) (*rg-set*) which is a superset of Miller and Charles's set.

3.1 Replication Attempts

This research emerged from a project running a similar experiment for Dutch on Cornetto (Vossen et al., 2013). First, an attempt was made to reproduce the results reported in Patwardhan and Pedersen (2006) and Pedersen (2010) on the English WordNet using their WordNet::Similarity web-interface.[3] Results differed from those reported in the aforementioned works, even when using the same versions as the original, WordNet::Similarity-1.02 and WordNet 2.1 (Patwardhan and Pedersen, 2006) and WordNet::Similarity-2.05 and WordNet 3.0 (Pedersen, 2010), respectively.[4]

The fact that results of similarity measures on WordNet can differ even while the same software and same versions are used indicates that properties which are not addressed in the literature may influence the output of similarity measures. We therefore conducted a range of experiments that, in addition to searching for the right settings to replicate results of previous research, address the following questions:

1) Which properties have an impact on the performance of WordNet similarity measures?

2) How much does the performance of individual measures vary?

3) How do commonly used measures compare when the variation of their performance are taken into account?

3.2 Methodology and first observations

The questions above were addressed in two stages. In the first stage, Fokkens, who was not involved in the first replication attempt implemented a script to calculate similarity measures using WordNet::Similarity. This included similarity measures introduced by Wu and Palmer (1994) (`wup`),

[3] Obtained from `http://talisker.d.umn.edu/cgi-bin/similarity/similarity.cgi`, WordNet::Similarity version 2.05. This web interface has now moved to `http://maraca.d.umn.edu`

[4] WordNet::Similarity were obtained `http://search.cpan.org/dist/WordNet-Similarity/`.

Leacock and Chodorow (1998) (lch), Resnik (1995) (res), Jiang and Conrath (1997) (jcn), Lin (1998) (lin), Banerjee and Pedersen (2003) (lesk), Hirst and St-Onge (1998) (hso) and Patwardhan and Pedersen (2006) (vector and vpairs) respectively.

Consequently, settings and properties were changed systematically and shared with Pedersen who attempted to produce the new results with his own implementations. First, we made sure that the script implemented by Fokkens could produce the same WordNet similarity scores for each individual word pair as those used to calculate the ranking on the *mc-set* by Pedersen (2010). Finally, the gold standard and exact implementation of the Spearman ranking coefficient were compared.

Differences in results turned out to be related to variations in the **experimental setup**. First, we made different assumptions on the restriction of part-of-speech tags (henceforth “PoS-tag”) considered in the comparison. Miller and Charles (1991) do not discuss how they deal with words with more than one PoS-tag in their study. Pedersen therefore included all senses with any PoS-tag in his study. The first replication attempt had restricted PoS-tags to nouns based on the idea that most items are nouns and subjects would be primed to primarily think of the noun senses. Both assumptions are reasonable. Pos-tags were not restricted in the second replication attempt, but because of a bug in the code only the first identified PoS-tag (“noun” in all cases) was considered. We therefore mistakenly assumed that PoS-tag restrictions did not matter until we compared individual scores between Pedersen and the replication attempts.

Second, there are two gold standards for the Miller and Charles (1991) set: one has the scores assigned during the original experiment run by Rubenstein and Goodenough (1965), the other has the scores assigned during Miller and Charles (1991)’s own experiment. The ranking correlation between the two sets is high, but they are not identical. Again, there is no reason why one gold standard would be a better choice than the other, but in order to replicate results, it must be known which of the two was used. Third, results changed because of differences in the treatment of ties while calculating Spearman ρ. The influence of the exact gold standard and calculation of Spearman ρ could only be found because Pedersen could provide the output of the similarity measures he used to calculate the coefficient. It is unlikely we would have been able to replicate his results at all without the output of this intermediate step. Finally, results for lch, lesk and wup changed according to measure specific configuration settings such as including a PoS-tag specific root node or turning on normalisation.

measure	Spearman ρ		Kendall τ		ranking
	min	max	min	max	variation
path based similarity					
path	0.70	0.78	0.55	0.62	1-8
wup	0.70	0.79	0.53	0.61	1-6
lch	0.70	0.78	0.55	0.62	1-7
path based information content					
res	0.65	0.75	0.26	0.57	4-11
lin	0.49	0.73	0.36	0.53	6-10
jcn	0.46	0.73	0.32	0.55	5, 7-11
path based relatedness					
hso	0.73	0.80	0.36	0.41	1-3,5-10
dictionary and corpus based relatedness					
vpairs	0.40	0.70	0.26	0.50	7-11
vector	0.48	0.92	0.33	0.76	1,2,4,6-11
lesk	0.66	0.83	-0.02	0.61	1-8,11,12

Table 1: Variation WordNet measures’ results

In the second stage of this research, we ran experiments that systematically manipulate the influential factors described above. In this experiment, we included both the *mc-set* and the complete *rg-set*. The implementation of Spearman ρ used in Pedersen (2010) assigned the lowest number in ranking to ties rather than the mean, resulting in an unjustified drop in results for scores that lead to many ties. We therefore experimented with a different correlation measure, Kendall tau coefficient (Kendall, 1938, τ) rather than two versions of Spearman ρ.

3.3 Variation per measure

All measures varied in their performance. The complete outcome of our experiments (both the similarity measures assigned to each pair as well as the output of the ranking coefficients) are included in the data set provided at http://github.com/antske/WordNetSimilarity. Table 1 presents an overview of the main point we wish to make through this experiment: the minimal and maximal results according to both ranking coefficients. Results for similarity measures varied from 0.06-0.42 points for Spearman ρ and from 0.05-0.60 points for Kendall τ. The last column indicates the variation of performance of a measure compared to the other measures, where 1 is the best performing measure and 12 is the worst.[5] For instance, `path` has been best performing measure, second best, eighth best and all positions in between, `vector` has ranked first, second and fourth, but also occupied all positions from six to eleven.

In principle, it is to be expected that numbers are not exactly the same while evaluating against a different data set (the *mc-set* versus the *rg-set*), taking a different set of synsets to evaluate on (changing PoS-tag restrictions) or changing configuration settings that influence the similarity score. However, a variation of up to 0.44 points in Spearman ρ and 0.60 in Kendall τ[6] leads to the question of how indicative these results really are. A more serious problem is the fact that the comparative performance of individual measure changes. Which measure performs best depends on the evaluation set, ranking coefficient, PoS-tag restrictions and configuration settings. This means that the answer to the question of which similarity measure is best to mimic human similarity scores depends on aspects that are often not even mentioned, let alone systematically compared.

3.4 Variation per category

For each influential category of experimental variation, we compared the variation in Spearman ρ and Kendall τ, while similarity measure and other influential categories were kept stable. The categories we varied include WordNet and WordNet::Similarity version, the gold standard used to evaluate, restrictions on PoS-tags, and measure specific configurations. Table 2 presents the maximum variation found across measures for each category. The last column indicates how often the ranking of a specific measure changed as the category changed, e.g. did the measure ranking third using specific configurations, PoS-tag restrictions and a specific gold standard using WordNet 2.1 still rank third when WordNet 3.0 was used instead? The number in parentheses next to the 'different ranks' in the table presents the total number of scores investigated. Note that this number changes for each category, because we compared two WordNet versions (WN version), three gold standard and PoS-tag restriction variations and configuration only for the subset of scores where configuration matters.

Variation	Maximum difference		Different
	Spearman ρ	Kendall τ	rank (tot)
WN version	0.44	0.42	223 (252)
gold standard	0.24	0.21	359 (504)
PoS-tag	0.09	0.08	208 (504)
configuration	0.08	0.60	37 (90)

Table 2: Variations per category

There are no definite statements to make as to which version (Patwardhan and Pedersen (2006) vs Pedersen (2010)), PoS-tag restriction or configuration gives the best results. Likewise, while most measures do better on the smaller data set, some achieve their highest results on the full set. This is partially due to the fact that ranking coefficients are sensitive to outliers. In several cases where PoS-tag restrictions led to different results, only one pair received a different score. For instance, `path` assigns a relatively high score to the pair *chord-smile* when verbs are included, because the hierarchy of verbs in WordNet is relatively flat. This effect is not observed in `wup` and `lch` which correct for the depth of the hierarchy. On the other hand, `res`, `lin` and `jcn` score better on the same set when verbs are considered, because they cannot detect any relatedness for the pair *crane-implement* when restricted to nouns.

On top of the variations presented above, we notice a discrepancy between the two coefficients. Kendall τ generally leads to lower coefficiency scores than Spearman ρ. Moreover, they each give different relative indications: where `lesk` achieves its highest Spearman ρ, it has an extremely low Kendall τ of 0.01. Spearman ρ uses the difference in rank as its basis to calculate a correlation, where Kendall τ uses the number of items with the correct rank. The low Kendall τ for `lesk` is the result of three pairs receiving a score that is too high. Other pairs that get a relatively accurate score are pushed one place down in rank. Because only items that receive the exact same rank help to increase τ, such a shift can result in a drastic drop in the coefficient. In our opinion, Spearman ρ is therefore preferable over Kendall τ. We included τ, because many authors do not mention the ranking coefficient they use (cf. Budanitsky and Hirst (2006), Resnik (1995)) and both ρ and τ are com-

[5] Some measures ranked differently as their individual configuration settings changed. In these cases, the measure was included in the overall ranking multiple times, which is why there are more ranking positions than measures.

[6] Section 3.4 explains why the variation in Kendall is this extreme and ρ is more appropriate for this task.

monly used coefficients.

Except for WordNet, which Budanitsky and Hirst (2006) hold accountable for minor variations in a footnote, the influential categories we investigated in this paper, to our knowledge, have not yet been addressed in the literature. Cramer (2008) points out that results from WordNet-Human similarity correlations lead to scattered results reporting variations similar to ours, but she compares studies using different measures, data and experimental setup. This study shows that even if the main properties are kept stable, results vary enough to change the identity of the measure that yields the best performance. Table 1 reveals a wide variation in ranking relative to alternative approaches. Results in Table 2 show that it is common for the ranking of a score to change due to variations that are not at the core of the method.

This study shows that it is far from clear how different WordNet similarity measures relate to each other. In fact, we do not know how we can obtain the best results. This is particularly challenging, because the 'best results' may depend on the intended use of the similarity scores (Meng et al., 2013). This is also the reason why we presented the maximum variation observed, rather than the average or typical variation (mostly below 0.10 points). The experiments presented in this paper resulted in a vast amount of data. An elaborate analysis of this data is needed to get a better understanding of how measures work and why results vary to such an extent. We leave this investigation to future work. If there is one take-home message from this experiment, it is that one should experiment with parameters such as restrictions on PoS-tags or configurations and determine which score to use depending on what it is used for, rather than picking something that did best in a study using different data for a different task and may have used a different version of WordNet.

4 Reproducing a NER method

Freire et al. (2012) describe an approach to classifying named entities in the cultural heritage domain. The approach is based on the assumption that domain knowledge, encoded in complex features, can aid a machine learning algorithm in NER tasks when only little training data is available. These features include information about person and organisation names, locations, as well as PoS-tags. Additionally, some general features are used such as a window of three preceding and two following tokens, token length and capitalisation information. Experiments are run in a 10-fold cross-validation setup using an open source machine learning toolkit (McCallum, 2002).

4.1 Reproducing NER Experiments

This experiment can be seen as a real-world case of *the sad tale of the Zigglebottom tagger* (Pedersen, 2008). The (fictional) Zigglebottom tagger is a tagger with spectacular results that looks like it will solve some major problems in your system. However, the code is not available and a new implementation does not yield the same results. The original authors cannot provide the necessary details to reproduce their results, because most of the work has been done by a PhD student who has finished and moved on to something else. In the end, the newly implemented Zigglebottom tagger is not used, because it does not lead to the promised better results and all effort went to waste.

Van Erp was interested in the NER approach presented in Freire et al. (2012). Unfortunately, the code could not be made available, so she decided to reimplement the approach. Despite feedback from Freire about particular details of the system, results remained 20 points below those reported in Freire et al. (2012) in overall F-score (Van Erp and Van der Meij, 2013).

The reimplementation process involved choices about seemingly small details such as rounding to how many decimals, how to tokenise or how much data cleanup to perform (normalisation of non-alphanumeric characters for example). Trying different parameter combinations for feature generation and the algorithm never yielded the exact same results as Freire et al. (2012). The results of the best run in our first reproduction attempt, together with the original results from Freire et al. (2012) are presented in Table 3. Van Erp and Van der Meij (2013) provide an overview of the implementation efforts.

4.2 Following up from reproduction

Since the experiments in Van Erp and Van der Meij (2013) introduce several new research questions regarding the influence of data cleaning and the limitations of the dataset, we performed some additional experiments.

First, we varied the tokenisation, removing non-alphanumeric characters from the data set. This yielded a significantly smaller data set (10,442

	(Freire et al., 2012) results			Van Erp and Van der Meij's replication results		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (388)	92%	55%	69	77.80%	39.18%	52.05
ORG (157)	90%	57%	70	65.75%	30.57%	41.74
PER (614)	91%	56%	69	73.33%	37.62%	49.73
Overall (1,159)	91%	55%	69	73.33%	37.19%	49.45

Table 3: Precision, recall and $F_{\beta=1}$ scores for the original experiments from Freire et al. 2012 and our replication of their approach as presented in Van Erp and Van der Meij (2013)

tokens vs 12,510), and a 15 point drop in overall F-score. Then, we investigated whether variation in the cross-validation splits made any difference as we noticed that some NEs were only present in particular fields in the data, which can have a significant impact on a small dataset. We inspected the difference between different cross-validation folds by computing the standard deviations of the scores and found deviations of up to 25 points in F-score between the 10 splits. In the general setup, database records were randomly distributed over the folds and cut off to balance the fold sizes. In a different approach to dividing the data by distributing individual sentences from the records over the folds, performance increases by 8.57 points in overall F-score to 58.02. This is not what was done in the original Freire et al. (2012) paper, but shows that the results obtained with this dataset are quite fragile.

As we worried about the complexity of the feature set relative to the size of the data set, we deviated somewhat from Freire et al. (2012)'s experiments in that we switched some features on and off. Removal of complex features pertaining to the window around the focus token improved our results by 3.84 points in overall F-score to 53.39. The complex features based on VIAF,[7] GeoNames[8] and WordNet do contribute to the classification in the Mallet setup as removing them and only using the focus token, window and generic features causes a slight drop in overall F-score from 49.45 to 47.25.

When training the Stanford NER system (Finkel et al., 2005) on just the tokens from the Freire data set and the parameters from english.all.3class.distsim.prop (included in the Stanford NER release, see also Van Erp and Van der Meij (2013)), our F-scores come very close to those reported by Freire et al. (2012), but mostly with a higher recall and lower precision. It is puzzling that the Stanford system obtains such high results with only very simple features, whereas for Mallet the complex features show improvement over simpler features. This leads to questions about the differences between the CRF implementations and the influence of their parameters, which we hope to investigate in future work.

4.3 Reproduction difficulties explained

Several reasons may be the cause of why we fail to reproduce results. As mentioned, not all resources and data were available for this experiment, thus causing us to navigate in the dark as we could not reverse-engineer intermediate steps, but only compare to the final precision, recall and F-scores.

The experiments follow a general machine learning setup consisting roughly of four steps: preprocess data, generate features, train model and test model. The novelty and replication problems lie in the first three steps. How the data was preprocessed is a major factor here. The data set consisted of XML files marked up with inline named entity tags. In order to generate machine learning features, this data has to be tokenised, possibly cleaned up and the named entity markup had to be converted to a token-based scheme. Each of these steps can be carried out in several ways, and choices made here can have great influence on the rest of the pipeline.

Similar choices have to be made for preprocessing external resources. From the descriptions in the original paper, it is unclear how records in VIAF and GeoNames were preprocessed, or even which versions of these resources were used. Preprocessing and calculating occurrence statistics over VIAF takes 30 hours for each run. It is thus not feasible to identify the main potential variations without the original data to verify this prepatory step.

Numbers had to be rounded when generating the features, leading to the question of how many decimals are required to be discriminative without creating an overly sparse dataset. Freire recalls that encoding features as multi-value discrete fea-

[7] `http://www.viaf.org`

[8] `http://www.geonames.org`

tures versus several boolean features can have significant impact. These settings are not mentioned in the paper, making reproduction very difficult.

As the project in which the original research was performed has ended, and there is no central repository where such information can be retrieved, we are left to wonder how to reuse this approach in order to further domain-specific NER.

5 Observations

In this section, we generalise the observations from our use cases to the main categories that can influence reproduction.

Despite our efforts to describe our systems as clearly as possible, details that can make a tremendous difference are often omitted in papers. It will be no surprise to researchers in the field that **preprocessing** of data can make or break an experiment.

The choice of which steps we perform, and how each of these steps is carried out exactly are part of our **experimental setup**. A major difference in the results for the NER experiments was caused by variations in the way in which we split the data for cross-validation.

As we fine-tune our techniques, software gets updated, data sets are extended or annotation bugs are fixed. In the WordNet experiment, we found that there were two different gold standard data sets. There are also different versions of WordNet, and the WordNet::Similarity packages. Similarly for the NER experiment, GeoNames, VIAF and Mallet are updated regularly. It is therefore critical to pay attention to **versioning**.

Our experiments often consist of several different steps whose outputs may be difficult to retrace. In order to check the output of a reproduction experiment at every step of the way, **system output** of experiments, including intermediate steps, is vital. The WordNet replication was only possible, because Pedersen could provide the similarity scores of each word pair. This enabled us to compare the intermediate output and identify the source of differences in output.

Lastly, there may be inherent **system variations** in the techniques used. Machine learning algorithms may for instance use coin flips in case of a tie. This was not observed in our experiments, but such variations may be determined by running an experiment several times and taking the average over the different runs (cf. Raeder et al. (2010)).

All together, these observations show that sharing data and software play a key role in gaining insight into how our methods work. Vanschoren et al. (2012) propose a setup that allows researchers to provide their full experimental setup, which should include exact steps followed in preprocessing the data, documentation of the experimental setup, exact versions of the software and resources used and experimental output. Having access to such a setup allows other researchers to validate research, but also tweak the approach to investigate system variation, systematically test the approach in order to learn its limitations and strengths and ultimately improve on it.

6 Discussion

Many of the aspects addressed in the previous section such as preprocessing are typically only mentioned in passing, or not at all. There is often not enough space to capture all details, and they are generally not the core of the research described. Still, our use cases have shown that they can have a tremendous impact on reproduction, and can even lead to different conclusions. This leads to serious questions on how we can interpret our results and how we can compare the performance of different methods. Is an improvement of a few per cent really due to the novelty of the approach if larger variations are found when the data is split differently? Is a method that does not quite achieve the highest reported state-of-the-art result truly less good? What does a state-of-the-art result mean if it is only tested on one data set?

If one really wants to know whether a result is better or worse than the state-of-the-art, the range of variation within the state-of-the-art must be known. Systematic experiments such as the ones we carried out for WordNet similarity and NER, can help determine this range. For results that fall within the range, it holds that they can only be judged by evaluations going beyond comparing performance numbers, i.e. an evaluation of how the approach achieves a given result and how that relates to alternative approaches.

Naturally, our use cases do not represent the entire gamut of research methodologies and problems in the NLP community. However, they do represent two core technologies and our observations align with previous literature on replication and reproduction.

Despite the systematic variation we employed in our experiments, they do not answer all questions that the problems in reproduction evoked. For the WordNet experiments, deeper analysis is required to gain full understanding of how individual influential aspects interact with each measurement. For the NER experiments, we are yet to identify the cause of our failure to reproduce.

The considerable time investment required for such experiments forms a challenge. Due to pressure to publish or other time limitations, they cannot be carried out for each evaluation. Therefore, it is important to share our experiments, so that other researchers (or students) can take this up. This could be stimulated by instituting reproduction tracks in conferences, thus rewarding systematic investigation of research approaches. It can also be aided by adopting initiatives that enable authors to easily include data, code and/or workflows with their publications such as the PLOS/figshare collaboration.[9] We already do a similar thing for our research problems by organising challenges or shared tasks, why not extend this to systematic testing of our approaches?

7 Conclusion

We have presented two reproduction use cases for the NLP domain. We show that repeating other researchers' experiments can lead to new research questions and provide new insights into and better understanding of the investigated techniques.

Our WordNet experiments show that the performance of similarity measures can be influenced by the PoS-tags considered, measure specific variations, the rank coefficient and the gold standard used for comparison. We not only find that such variations lead to different numbers, but also different rankings of the individual measures, i.e. these aspects lead to a different answer to the question as to which measure performs best. We did not succeed in reproducing the NER results of Freire et al. (2012), showing the complexity of what seems a straightforward reproduction case based on a system description and training data only. Our analyses show that it is still an open question whether additional complex features improve domain specific NER and that this may partially depend on the CRF implementation.

Some observations go beyond our use cases. In particular, the fact that results vary significantly because of details that are not made explicit in our publications. Systematic testing can provide an indication of this variation. We have classified relevant aspects in five categories occurring across subdisciplines of NLP: **preprocessing**, **experimental setup**, **versioning**, **system output**, and **system variation**.

We believe that knowing the influence of different aspects in our experimental workflow can help increase our understanding of the robustness of the approach at hand and will help understand the meaning of the state-of-the-art better. Some techniques are reused so often (the papers introducing WordNet similarity measures have around 1,000-2,000 citations each as of February 2013, for example) that knowing their strengths and weaknesses is essential for optimising their use.

As mentioned many times before, sharing is key to facilitating reuse, even if the code is imperfect and contains hacks and possibly bugs. In the end, the same holds for software as for documentation: *it is like sex: if it is good, it is very good and if it is bad, it is better than nothing!*[10] But most of all: when reproduction fails, regardless of whether original code or a reimplementation was used, valuable insights can emerge from investigating the cause of this failure. So don't let your failing reimplementations of the *Zigglebottom* tagger collect dusk on a shelf while others reimplement their own failing *Zigglebottoms*. As a community, we need to know where our approaches fail, as much –if not more– as where they succeed.

Acknowledgments

We would like to thank the anonymous reviewers for their eye to detail and useful comments to make this a better paper. We furthermore thank Ruben Izquierdo, Lourens van der Meij, Christoph Zwirello, Rebecca Dridan and the Semantic Web Group at VU University for their help and useful feedback. The research leading to this paper was supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404), the Agora project, by NWO CATCH programme, grant 640.004.801, and the BiographyNed project, a joint project with Huygens/ING Institute of the Dutch Academy of Sciences funded by the Netherlands eScience Center (`http://esciencecenter.nl/`).

[9] `http://blogs.plos.org/plos/2013/01/easier-access-to-plos-data/`

[10] The documentation variant of this quote is attributed to Dick Brandon.

References

Stanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August.

Daniel M. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Tomasz Buchert and Lucas Nussbaum. 2012. Leveraging business workflows in distributed systems research for the orchestration of reproducible and scalable experiments. In Anne Etien, editor, *9ème édition de la conférence MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication - MajecSTIC 2012 (2012)*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Phd dissertation, University of Pennsylvania.

Irene Cramer. 2008. How well do semantic relatedness measures perform? a meta-study. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1, pages 59–70.

Olivier Dalle. 2012. On reproducibility and traceability of simulations. In *WSC-Winter Simulation Conference-2012*.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning IV*.

Jenny Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, USA.

Nuno Freire, José Borbinha, and Pável Calado. 2012. An approach for named entity recognition in poorly structured data. In *Proceedings of ESWC 2012*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. MIT Press.

James Howison and James D. Herbsleb. 2013. Sharing the spoils: incentives and collaboration in scientific software development. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work*, pages 459–470.

Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. 2012. The case for open computer programs. *Nature*, 482(7386):485–488.

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, Taiwan.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

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

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

Panos Louridas and Georgios Gousios. 2012. A note on rigour and replicability. *SIGSOFT Softw. Eng. Notes*, 37(5):1–4.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Thilo Mende. 2010. Replication of defect prediction studies: problems, pitfalls and recommendations. In *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*. ACM.

Lingling Meng, Runqing Huang, and Junzhong Gu. 2013. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12.

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

Cameron Neylon, Jan Aerts, C Titus Brown, Simon J Coles, Les Hatton, Daniel Lemire, K Jarrod Millman, Peter Murray-Rust, Fernando Perez, Neil Saunders, Nigam Shah, Arfon Smith, Gaël Varoquaux, and Egon Willighagen. 2012. Changing computational research. the challenges ahead. *Source Code for Biology and Medicine*, 7(2).

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*, pages 1–8, Trento, Italy.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

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)*, pages 329–332, Los Angeles, USA.

Troy Raeder, T. Ryan Hoens, and Nitesh V. Chawla. 2010. Consequences of variability in classifier performance estimates. In *Proceedings of ICDM'2010*.

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)*, pages 448–453, Montreal, Canada.

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

Wheeler Ruml. 2010. The logic of benchmarking: A case against state-of-the-art performance. In *Proceedings of the Third Annual Symposium on Combinatorial Search (SOCS-10)*.

Charles Spearman. 1904. Proof and measurement of association between two things. *American Journal of Psychology*, 15:72—101.

Marieke Van Erp and Lourens Van der Meij. 2013. Reusable research? a case study in named entity recognition. CLTL 2013-01, Computational Lexicology & Terminology Lab, VU University Amsterdam.

Joaquin Vanschoren, Hendrik Blockeel, Bernhard Pfahringer, and Geoffrey Holmes. 2012. Experiment databases. *Machine Learning*, 87(2):127–158.

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 Combinatorial Lexical Semantic Database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch Results by the STEVIN-programme*, number XVII in Theory and Applications of Natural Language Processing, chapter 10. Springer.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133—138, Las Cruces, USA.