Finding Synonyms in Medical Texts
– Creating a system for automatic synonym extraction from medical texts

Gustav Cederblad

Supervisor: Arne Jönsson
Assistant supervisor: Marina Santini
Examiner: Henrik Danielsson
Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Gustav Cederblad
Abstract

This thesis describes the work of creating an automatic system for identifying synonyms and semantically related words in medical texts. Before this work, as a part of the project E-care@home, medical texts have been classified as either lay or specialized by both a lay annotator and an expert annotator. The lay annotator, in this case, is a person without any medical knowledge, whereas the expert annotator has professional knowledge in medicine. Using these texts made it possible to create co-occurrences matrices from which the related words could be identified. Fifteen medical terms were chosen as system input. The Dice similarity of these words in a context window of ten words around them was calculated. As output, five candidate related terms for each medical term was returned. Only unigrams were considered. The candidate related terms were evaluated using a questionnaire, where 223 healthcare professionals rated the similarity using a scale from one to five. A Fleiss kappa test showed that the agreement among these raters was 0.28, which is a fair agreement. The evaluation further showed that there was a significant correlation between the human ratings and the relatedness score (Dice similarity). That is, words with higher Dice similarity tended to get a higher human rating. However, the Dice similarity interval in which the words got the highest average human rating was 0.35-0.39. This result means that there is much room for improving the system. Further developments of the system should remove the unigram limitation and expand the corpus to provide a more accurate and reliable result.

Keywords: eHealth, distributional semantics, medical synonyms, semantic relations, word similarity
Acknowledgments

I would like to thank my supervisor Arne Jönsson for supporting me along the way.

A special thanks to my assistant supervisor Marina Santini, who no matter the question, always answered positively and encouraged me to keep on working and exploring.

To my sister who provided great help when distributing my questionnaire and for sharing her vast knowledge of medical words.
# Table of contents

1 Introduction ................................................................................................................. 1

1.1 About the project .................................................................................................... 1

1.2 Purpose .................................................................................................................... 1

1.3 Delimitation ............................................................................................................. 1

2 Theory ....................................................................................................................... 3

2.1 Semantic relationships ............................................................................................ 3

2.2 Distributional semantics ......................................................................................... 3

2.3 Sparse matrices ....................................................................................................... 4

2.4 Weighting ................................................................................................................ 4

2.5 Similarity metrics .................................................................................................. 5

2.6 Previous work ......................................................................................................... 7

2.7 Synonym list or thesaurus ....................................................................................... 9

3 Method ...................................................................................................................... 10

3.1 Data ....................................................................................................................... 10

3.2 Pre-processing ....................................................................................................... 10

3.3 Implementation ...................................................................................................... 11

3.4 Finding semantically related words ....................................................................... 14

3.5 Evaluation ............................................................................................................. 15

4 Results ...................................................................................................................... 20

4.1 System output ....................................................................................................... 20

4.2 Evaluation results ................................................................................................. 21

5 Discussion ............................................................................................................... 25

5.1 Main results ........................................................................................................... 25

5.2 Limited corpus ...................................................................................................... 25

5.3 Performance intervals ......................................................................................... 26

5.4 Unigram limitation ............................................................................................... 26

5.5 Noise ..................................................................................................................... 27

5.6 Evaluation method .............................................................................................. 27

5.7 Lay and expert difference ..................................................................................... 28

5.8 Real-world usage ................................................................................................. 28

5.9 Further research .................................................................................................. 28

6 Conclusion ............................................................................................................... 31

References .................................................................................................................. 32
List of Figures

Figure 1. Graphical description of where in the healthcare chain this project might be used... 1
Figure 2. A description of the procedure of extracting the candidate related terms. ............ 15
Figure 3. The performance of the system at specific intervals......................................... 23
List of Tables

Table 1. An example of a word-word matrix containing made-up data............................... 6
Table 2. The comparison of similarity measures............................................................... 14
Table 3. Landis and Koch (1977) proposal of Kappa interpretation.................................. 18
Table 4. The system output, showing each SNOMED term’s five most similar words as well as the Dice similarity................................................................. 20
Table 5. The size of the extracted context windows......................................................... 21
Table 6. The agreement rate of each category................................................................. 22
Table 7. The highest user rated candidate related term for every SNOMED term. ............ 22
Table 8. The values used in Figure 3.................................................................................. 23
Table 9. Percentage of “I don’t know” answers. .............................................................. 24
1 Introduction

The following chapter describes the project to which the current work aims contribute, as well as the purpose of the current work and its delimitations.

1.1 About the project

The work described in this thesis is done in collaboration with E-care@home, which is a Swedish project aiming to support the care of older adults in their homes. Placing different sensors in a person’s home makes it possible to collect much useful information. This information can then be sent to caregivers and other persons responsible for the health of an older adult. There could, however, be a knowledge gap between the healthcare professionals and the older adult, salient in their use of language, which may lead to misunderstandings and miscommunication. Therefore, making it easier for patients to take part of information about themselves is important (Santini, Jönsson, Nyström, & Alirezai, 2017).

1.2 Purpose

The purpose of the current project is to extract synonyms and related words from a corpus consisting of medical texts and to evaluate whether these words actually are related. Extracting these medical synonyms could make it easier for laypeople to understand the medical language that often can be quite complicated. Swedish words that may be well known to the public can often have a Latin synonym that is more commonly used by experts. Sometimes, the more specific medical words can be described by one more broad and straightforward word that is more familiar to lay people. If the synonyms are identified, words that are only being used by experts can be changed to their lay synonyms, thus improving the readability and thereby bridging the knowledge gap between lay people and experts. Figure 1 describes where the focus of this project lies and wherein the communication between patients and healthcare professionals the system could be implemented.

![Figure 1](image.png)

*Figure 1. Graphical description of where in the healthcare chain this project might be used.*

1.3 Delimitation

For this project, the aim was only to find synonyms and related words for different diseases. Only unigrams were considered. Thus, medical words of parts of speech other than nouns were excluded since synonyms always belong to the same part of speech. This exclusion was to limit the extent of the analysis and evaluation.
The corpus used was constructed of texts publicly available on the internet. No texts from actual medical records were used, although this is where the project aims to be applied.

This project did not evaluate whether the extracted words were more comprehensible to lay people. An evaluation of this kind must be conducted in further studies.

The processing was done mostly on already collected and annotated corpora. Thus, a significant part of the project was focused on gathering relevant literature and writing code for processing the corpora.
2 Theory

Starting with the definitions of semantic relationships, this chapter will describe the theoretical ground used for accomplishing the task of identifying synonyms and related words in an automatic way. Presented in this chapter are methods for creating co-occurrence matrices, weighting, and calculating similarity. Furthermore, some work has already been done on the same topic, and these studies are briefly described. In the remainder of the thesis, some Swedish words will be used in examples. For their English equivalent, please consult Appendix E.

2.1 Semantic relationships

In a language, there are several different types of semantic relationships. Therefore, it can be useful to state their definitions explicitly. This work aimed to find synonyms, but words with other semantical relations also had to be included to widen the scope.

**Synonyms** – words that share a similar meaning. It is important to emphasize the “similar meaning” since words with the same meaning, called perfect synonyms, are seldom found. Some synonyms may come with positive or negative connotations, although their meaning is similar (Saeed, 2015). In Swedish medical language, “blåmärke” and “hematom” are synonyms. “Hematom” means small internal bleeding\(^1\), and by this definition, it comprises several other terms. Only if trauma causes the bleeding, it is called “blåmärke”\(^2\). Thus, “hematom” is always a synonym to “blåmärke,” but blåmärke is not always a synonym to “hematom.” These kinds of small differences are common in the case of synonymy, and this is an excellent example of why perfect synonyms are rare.

**Antonyms** – words that have the opposite meaning to each other (Saeed, 2015).

**Hyponyms** – words that are included in a broader kind of word (*car* and *bike* are hyponyms of *vehicle*) (Saeed, 2015).

**Hypernyms** – words that include other more specific words (in the example above, the word *vehicle* is the hypernym of *car* and *bike*) (Saeed, 2015).

2.2 Distributional semantics

Language is not considered to be structured in an arbitrary manner. On the contrary, words related to each other often occur in a similar context (Harris, 1954). For example, apple and pear should be surrounded by similar words in any given context, since they are hyponyms. The two words are both describing a kind of fruit and are both edible. Thus, it is likely that they will appear next to words regarding fruits or eating.

Words can be represented as vectors in a high-dimensional space where the location of the vectors are dependent on the words’ distribution, and thereby their meaning. This is what is called a word-space or semantic-space model (Lund & Burgess, 1996; Sahlgren, 2006). Representing words as vectors enable linear algebra operations to be performed in a semantic context, making it possible to calculate the similarity between words (Jurafsky & Martin, 2017). The closer two vectors are in this high-dimensional space, the more semantically similar they are. This calculation is done without even having to know the semantic meaning of the words. Using

\(^1\) According to the definition at: https://mesh.kib.ki.se/term/D006406/hematoma

\(^2\) According to the definition at: https://mesh.kib.ki.se/term/D003288/contusions
simple word-space models, determining the exact semantic relationship between words might be difficult. That is, whether they, for example, are synonyms or antonyms (Sahlgren, 2006).

The word vectors are based on a co-occurrence matrix where each dimension represents the frequency of a word’s occurrence in either a document or together with another word. There are two different types of matrices for this purpose, a term-document matrix or a word-word matrix and they are further described below.

**2.2.1 Term-document matrix**
As the name implies, the term-document matrix consists of terms and the document to which each term belongs. Every word in the vocabulary is a row, and the columns are every document. The numbers in the matrix represent the occurrence of a word in a document. Then, a word could be represented as a vector, where each document represents a dimension (Jurafsky & Martin, 2017).

**2.2.2 Term-context matrix**
Instead of showing how often a word occurs in a document, the term-context matrix shows how often two words occur in the same context. This kind of matrix is also called word-word matrix. Thus, both the rows and the columns correspond to a particular word, and a word vector is constructed from the word’s occurrences together with other words (Jurafsky & Martin, 2017).

**2.3 Sparse matrices**
The dimensionality of the matrices mentioned above can quickly become quite large. The more data it contains, the more laborious analyzing it gets. If one were to create a word-word matrix containing 4000 unique words, the matrix would consist of $4000 \times 4000 = 16,000,000$ values. The growth is exponential; thus, it is easy to see why methods for reducing the number of dimensions are needed.

In most cases, a significant portion of a co-occurrence matrix will be zeros. In the example above, there is a vocabulary of 4000 words. This means that in a term-document matrix, the occurrence of all these 4000 words will be counted. Naturally, most of these words will never occur in most of the documents. Usually, 99% of the matrix consists of zeros (Sahlgren, 2005). This phenomenon is what is called sparsity – it is a sparse matrix. There are several methods for handling sparse matrices, such as Random Indexing (RI) and Latent Semantic Analysis (LSA). However, none of these methods was used in the current work.

**2.4 Weighting**
More than 60 years ago, the British linguist John Rupert Firth stated that words are to be known by the company they keep. This statement is quite relevant in this very context. Just counting the occurrences of a word in a particular context can tell us a lot, but more importantly, words occurring in the same context, close to the target word, could give us a hint about the word’s true meaning (Jurafsky & Martin, 2017). This is where weighting comes into the picture. Two different weighting methods will be mentioned in this section; PMI and Tf-idf.

**2.4.1 Pointwise mutual information (PMI)**
Pointwise mutual information (PMI) is a measure that takes into consideration how often two words occur in the same context (Jurafsky & Martin, 2017).

$$PMI(x, y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$ (1)
The idea behind this measurement is to consider whether the probability of two words appearing together within the same context is higher than that of the probability of them appearing in the same context by chance. Thus, the formula, which can be seen in Equation 1, for calculating this measurement is the probability of the words \( x \) and \( y \) occurring together, \( P(x,y) \), divided by the probability of the words occurring separately (Jurafsky & Martin, 2017). From this notion, one can intuitively see that if the probability of the words appearing together is higher than that of them appearing individually, the PMI value is higher.

The founding father of PMI, or mutual information as it was called at the time, is the computer scientist Robert Fano (Jurafsky & Martin, 2017). Church and Hanks (1989) developed it further by using a specific window of words. It was stated that smaller window sizes were beneficial when trying to identify expressions where the words are fixed. Fixed expressions can, for example, be idioms. Furthermore, they state that deciding on a general window size is difficult since it is dependent on the context (Church & Hanks, 1989).

A variant of PMI is Positive PMI (PPMI). The difference here is that all negative values are considered as a 0. This method is often used because negative values are considered unreliable (Jurafsky & Martin, 2017).

### 2.4.2 Tf-idf

Tf-idf (Term frequency – inverse document frequency) is another weighting method for co-occurrence matrices. Just as in the case of PMI, some terms contain more information or can say more about the context than others. With Tf-idf, rare words are weighted higher than common words. The value is calculated by dividing the number of documents by the total occurrence of the current word. This quota is multiplied by the occurrence of the current word in a certain document (Jurafsky & Martin, 2017).

### 2.5 Similarity metrics

When a co-occurrence matrix has been created, and the occurrences have been weighted, what is left is calculating the similarity. For this calculation, there is a wide range of different methods, but the focus of the current work has been on Cosine, Jaccard and Dice similarity.

#### 2.5.1 Cosine similarity

Sahlgren (2006) explains that one common way to measure the similarity between two vectors – in this case representing two words – is to calculate the cosine value of the angle between two vectors. The maximum dissimilarity between two vectors is when the angle is 180 degrees, resulting in a cosine similarity of -1. Since co-occurrence matrices only contain values above zero, the cosine similarity will never be negative.

The cosine similarity is calculated using the dot product of two vectors.

\[
a \cdot b = \|a\| \cdot \|b\| \cos \theta (2)
\]

When calculating the dot product of the vectors \( a \) and \( b \), the length of the vectors (\( \|a\| \) and \( \|b\| \)) are multiplied with the cosine of the angle between them.

\[
\cos \theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} (3)
\]

Dividing the product of the two vectors with the products of their length, as seen in Equation 3, returns the cosine value of the angle, which is their cosine similarity. This division is done to
normalize the vectors. Otherwise, the results would be skewed because longer vectors would be counted as similar to most other vectors. That is, in a word space, more frequently occurring words would be counted as similar to many other words since increased length would increase the cosine similarity (Sahlgren, 2006).

Table 1. An example of a word-word matrix containing made-up data.

<table>
<thead>
<tr>
<th></th>
<th>lunga</th>
<th>klase</th>
<th>lungblåsa</th>
<th>lungemfysem</th>
<th>rökning</th>
</tr>
</thead>
<tbody>
<tr>
<td>lunga</td>
<td>9</td>
<td>8</td>
<td>19</td>
<td>41</td>
<td>9</td>
</tr>
<tr>
<td>klase</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>lungblåsa</td>
<td>19</td>
<td>3</td>
<td>11</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>lungemfysem</td>
<td>41</td>
<td>2</td>
<td>23</td>
<td>37</td>
<td>17</td>
</tr>
<tr>
<td>rökning</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>17</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1 shows a simple co-occurrence matrix, in this case, a word-word matrix. The data is made-up to provide an example of how to calculate the similarity between two words. Two rows are marked with red. These rows are examples of what constitutes word vectors with five dimensions. Using the word vectors as input in the Equation 3 will result in an equation like the one seen in Equation 4.

\[
\cos \theta = \frac{9 \cdot 41 + 8 \cdot 2 + 19 \cdot 23 + 41 \cdot 37 + 9 \cdot 17}{\sqrt{9^2 + 8^2 + 19^2 + 41^2 + 37^2 + 9^2}} = \frac{2492}{2963} \approx 0.84 \ (4)
\]

The conclusion from the previous calculation is that the words “lunga” and “lungemfysem” have a cosine similarity of 0.84 – a rather high similarity. Observable in the matrix is that the first dimension is the one that differs the most. Changing the first dimension for “lungemfysem” from 41 to 9 would result in a cosine similarity of 0.97. That is, vectors with similar co-occurrence frequency in the same dimensions tend to have a high similarity, which is in line with what Harris (1954) described.

2.5.2 Jaccard and Dice similarity

Two other similarity measurements are Jaccard and Dice, which have much in common. The Jaccard similarity is calculated using the formula in Equation 5.

\[
Sim_{\text{Jaccard}}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} \max(v_i, w_i)} \quad (5)
\]

This way of calculating similarity takes use of the intersection or overlap of two vectors. That is why the minimum value of the two vectors in each dimension is summed in the numerator, whereas the denominator is used for normalization (Jurafsky & Martin, 2017). By using the sample data from the previous section, the method for calculating the Jaccard similarity would produce the result in Equation 6.

\[
Sim_{\text{Jaccard}}(\text{lunga}, \text{lungemfysem}) = \frac{9+2+19+37+9}{41+8+23+41+17} = \frac{76}{130} \approx 0.58 \ (6)
\]

The Dice similarity is pretty similar to Jaccard, with the difference of a numerator multiplied by a factor of two, and with the sum of all dimensions as the denominator. This formula can be seen in Equation 7.
Calculating the Dice similarity on the same data as for Cosine and Jaccard produces the following result in Equation 8.

\[
Sim_{\text{Dice}}(\vec{v}, \vec{w}) = \frac{2 \times \sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} (v_i + w_i)} \tag{7}
\]

There is no use in comparing the values from the different methods. Only their internal difference matters. However, what these methods share is that their values span between zero and one, where a similarity of one implies that two vectors are the same.

### 2.6 Previous work

Simplifying text by replacing synonyms has been done before. Keski-Sarkkä and Jönsson (2012) used a synonym lexicon for finding the synonyms to replace in texts from four different genres. One of the things they conclude is that replacing synonyms is not as simple as it may seem. The choice of a word often depends on the context, and therefore, just replacing words may not always work. Based on their method for replacing words and evaluating the system, Abrahamsson, Forni, Skeppstedt, and Kivist (2014) worked with replacing synonyms in a medical context. For assessing readability, two commonly used measurements are LIX (läsbarhetsindex, Eng: readability measure) and OVIX (ordvariationsindex, Eng: word-variation index). The LIX value takes into consideration the number of words and sentences, as well as the number of long words. The longer a sentence is, and the longer words a text contains, the less readable it is. OVIX, on the other hand, uses the number of unique word in a text, to measure the lexical variance. Texts containing more unique words are considered more difficult to read. The results show that replacing synonyms in a medical text decreases the readability according to the LIX value while the readability increases according to the OVIX value. A possible explanation for this, according to Abrahamsson et al. (2014), is that more advanced Greek and Latin words have been replaced with longer yet simpler words, resulting in a higher LIX value. In nearly a third of the sentences that had words replaced, the meaning of the sentences changed. This is a problem that is likely to occur in the current work and thus has to be considered upon analyzing the results. Both the studies mentioned above used readability measures for evaluating their systems. However, in the current work, the aim is only to identify the synonyms, and therefore, evaluations like these are not feasible.

Henriksson et al. (2012) conducted a study, investigating how different combinations of random indexing and random permutation could be used to identify synonyms in medical texts. Random Indexing (RI) is a method that creates a co-occurrence matrix by first assigning a distributional vector, consisting of zeroes of a predetermined number of dimensions, to each word. Additionally, each word gets an index vector consisting of randomly distributed non-zero elements; -1 and 1. If one word occurs next to another, its index vector is added to the neighbor’s distributional vector. Random Permutation (RP) is very similar to RI, but also takes into consideration the order in which the words come. In addition to finding synonyms, they looked at abbreviation-expansion pairs, in both directions. Five different experimental setups were used; random indexing, random permutation, random indexing with random permutation filtering, random permutation with random indexing filtering, and random indexing together with random per-
mutation. Furthermore, they tested different window sizes. Their results showed that a combination of random indexing and random permutation, with window sizes of 20 respectively four, accomplished the best result. This setup was the best when the task was to identify synonyms. The conclusions from their study was that when using both RI and RP, the model takes into consideration two different aspects of a word: both its distributional and grammatical properties. Thus, they ought to be more similar than when only one aspect is considered (Henriksson et al., 2012).

Two years after the previously mentioned article, the same authors published a report on the same subject. Once again, a combination of random indexing and random permutation proved to be highly efficient in completing the task. The random indexing model used 1000 dimensions, and the index vectors consisted of eight non-zero elements. In this study, they limited it only to include words occurring more than 50 times. This limitation is motivated by the fact that the fewer times a word occurs in a corpus, the more difficult it is to model its meaning. The results also showed that the use of different corpora facilitated a better performance. Two different corpora were used: one consisting of Swedish health records and one composed of the issues of a Swedish medical journal. The reason for the combination of random indexing and random permutation being better is because they differ in the way that they model semantic relations (Henriksson, Moen, Skeppstedt, Daudaravičius, & Duneld, 2014).

Kann and Rosell (2005) used crowdsourcing for creating a Swedish dictionary of synonyms. After putting together a list of possible synonyms, they let people judge whether two words were synonyms. The system was implemented to the online service Lexin on-line, which is a Swedish lexicon. There was no need to define the semantic relationship that synonymy entails. On the contrary, people’s perception of synonymy is what mattered (Kann, 2004). Whenever a person looked something up on the Lexin website, two words were shown, asking the user to rate the synonymy on a scale from zero to five, with zero representing no synonymy and five representing full synonymy. There was also an “I don’t know” option. When enough people had rated a synonym, the system could decide to either delete the synonym or keep it, depending on the average rating. Crowdsourcing proved to be a useful tool in this case. That is, the system received more than two million gradings in five months. One of the lessons learned from this project is that presenting word pairs that are bad (far from synonyms) might make the users irritated (Kann & Rosell, 2005). This quality issue is something that must be considered when presenting the list of possible synonyms.

Before constructing an automatic system for synonym identification, it can be useful to know what characterizes medical language, and what distinguishes it from general language. Smith, Megyesi, Velupillai, and Kvist (2014) investigated this aspect of Swedish medical language. What their results show is that, in medical texts, abbreviations and more technical terms occur more frequently than in general language. Their results also show that there is a difference between texts from different parts of the healthcare system. Paying attention to abbreviations will be important in the current work, since they may hide an expansion of several words when the system only has the ability to handle unigrams.

2.6.1 MeSH
In prior research by Henriksson et al. (2012), the synonyms were gathered from Medical Subject Headings (MeSH). MeSH is a medical database provided by the U.S. National Library of Medicine. The terms in MeSH are in English. Therefore, the library at Karolinska Institutet is
working with translating it into a Swedish version. The MeSH database provides synonyms for many Swedish medical terms, although it does not contain synonyms for all the terms used in the current work.

2.7 Synonym list or thesaurus

The definition of a synonym list should be as Saeed (2015) mentions, words that share a similar meaning. Although, the words of interest in this work is not only synonyms by this definition. The question here is what one should call a list of words that are related to each other, but not necessarily share the same meaning. This gap is wherein the term thesaurus comes in handy.

A thesaurus or thesauri as it is also called is a list of words with a similar meaning. However, it also includes words that are semantically related, such as hypernyms and hyponyms (Sahlgren, 2006). Sahlgren (2006) further states that a thesaurus contains many arbitrary terms, making it difficult to create using a word-space model. An example of this arbitrariness is easily found when looking at a thesaurus entry for the word “demon,” where two of its supposedly related words are “enthusiast” and “visionary.” These words’ relation to each other is far-fetched.

A thesaurus, rather than a synonym list, is what this work aims to result in, allowing more words to be included. Although it might not be possible to replace a medical word with all the words in its corresponding thesaurus entry, a thesaurus may give the reader a sense of the word’s meaning.
3 Method

The method used in the current work consists of three major parts, pre-processing of the data, implementation of the system, and evaluation of the system output. Each of these parts will be extensively described in the following chapter.

3.1 Data

In the E-care@home project, there was a need for an annotated medical corpus. Santini et al. (2017) therefore worked with creating this. First, 1300 terms from the Swedish version of SNOMED CT was picked. The Swedish version of SNOMED (Systematized Nomenclature of Medicine) is a medical terminology provided by Socialstyrelsen. SNOMED aims to maintain consistent use of terms through different healthcare systems (Socialstyrelsen, n.d.). Out of these 1300 terms, some were considered irrelevant, resulting in using only 228 terms. The extracted terms were used as search queries in Google, and the resulting 843 URLs returned by Google were used in the BootCat software, for collecting the desired documents. All the documents were then annotated by a lay annotator (Santini et al., 2017).

The data used in this work is a further developed version of the corpus mentioned above. This corpus consists of medical texts on two different language levels, also denoted as sublanguages; lay and specialized. If a text is annotated as lay, it means that it is written on a level comprehensible to a person without any medical knowledge. The text annotated as specialized on the other hand is written on a level comprehensible only to a person with medical knowledge.

Each text has been annotated two times, both by a layperson and an expert. A layperson is a person without any medical knowledge, whereas the expert annotator is a person with general knowledge of medicine.

3.2 Pre-processing

For this project, much of the information stored in the corpus is unnecessary. For example, information about which website each text is crawled from is irrelevant. The first step of processing the data is, therefore, to get rid of this information, leaving only the necessary information.

As the corpus consists of several XML files, the tags can be used to gather only the relevant information. The structure of the XML files is shown in Appendix A. In this project, what is relevant is the text, which is stored under savedText.

3.2.1 Noise

Crawling text from the web is an effective way of gathering large pieces of text. However, this method might result in a somewhat noisy text. That is, it might contain many misspelled words, unknown characters, and URLs. If the noise is included in the later analysis, it will probably negatively affect the results.

Regular expressions are a great way of filtering out irrelevant data. Although, the use of regular expressions presupposes some knowledge of the texts’ structure. That is, it is necessary to know what characters to keep, change, or remove. An example of this is punctuation marks, which are easily removed and replaced with nothing. It is reasonable to assume that, in standard text, there is a space after all punctuation marks. The eCare corpus was not following these common rules – creating problems. As a result, additional code for handling these exceptions was
needed. Otherwise, these words were concatenated from, for example, “normal.Normalt” into “normalNormalt.” The resulting problem is obvious.

The approach to cleaning the corpus by finding what characters to remove was as follows. First, visually inspecting the corpus, finding what was visible. Secondly, using the print function in Python, which will return a UnicodeDecodeError whenever a non-Unicode character appears. This made it possible to identify these characters and wherein the corpus they occurred. Thirdly, using the corpus as input in R. By doing this, the R script could read all text until it encountered a non-Unicode character. The few non-Unicode characters left at this stage were manually removed when an error occurred in R.

### 3.3 Implementation

The processing of the corpora was done in R, using the “tm” and “quanteda” packages for text mining and matrix creation. The advantage of using these packages is that they have many useful built-in functions, such as functions for removing stop words, punctuation and extracting context windows.

#### 3.3.1 The “tm” package

The “tm” package was used to perform some of the pre-processing. In the E-care@home project, there were some already created R scripts for pre-processing the corpus. These scripts made use of the tm package when filtering out irrelevant data. Moreover, one crucial part was one script’s ability to split the corpus into one lay and one specialized.

#### 3.3.2 The “quanteda” package

The quanteda package is a brand new package for text processing in R. Some great advantages are the possibility to create document-feature matrices (dfm) and feature-context matrices (fcm). As input, the fcm function can take a corpus as well as tokens (Benoit et al., 2018). In the current study, the input was context windows. This package has a built-in list of Swedish stop words, although this list proved to be insufficient. Therefore, this list was extended with an extensive list taken from GitHub3. All stop words were stored in a text file which was loaded every time the program was run. In addition to these stop words, all words with a non-desirable part-of-speech tag were added to this list, which provided a convenient way of removing words from the analysis.

It was decided not to create a full co-occurrence matrix, to make the data more manageable. A paradoxical problem occurred at this stage. The problem was that when trying to use built-in functions for RI or LSA, there were too little available data. Although, creating a full co-occurrence matrix proved to be too computationally heavy, needing to much RAM to be able to run.

Instead, only context windows were extracted for the desired words. For this extraction, there is a built-in function in the “quanteda” package, namely KWIC (Key Word In Context). This function provides the possibility to input text, keyword, and window size, getting the desired windows as output. Using the KWIC function, it was possible to build a text file, with each line corresponding to an extracted context window. Setting the window size to ten, will, if possible, return a line of 21 words, with the desired word in the middle and the then words to its left and right. These extracted windows are then used as input when creating the fcm.

---

3 https://github.com/stopwords-iso/stopwords-sv
3.3.3 The textstat_simil function

The previously described quanteda package also has functions for calculating similarity, namely the textstat_simil function that can calculate a variety of different similarities. The function takes a word and returns a list of the most similar words. How many words the list contains is customizable.

In default mode, textstat_simil does not apply any weighting scheme. Only the raw frequencies are used. The problem with this is that the function does not take into consideration how close, in relation to each other, the words occur in the text. A frequently occurring word in the text is likely to be counted as more similar, than a less frequently occurring word, although this word is more closely related to the input word. Overcoming this obstacle is done by changing the weighting scheme. For this, there are two possibilities; the dfm_weight() function, or the dfm_tdidf() function. The dfm_weight() function has the following set of weighting schemes: "count", "prop", "propmax", "logcount", "boolean", "augmented", "logave", whereas the dfm_tdidf() function, as the name implies, uses Tf-idf weighting which is mentioned in section 2.4.2.

3.3.4 Part-of-speech filtering

During the first trails of the implemented model, it was possible to observe some words that, taken out of context, had no apparent relation to the target word. Considering this, a method for systematically removing those unrelated words was needed.

The words chosen for the evaluation are words from SNOMED describing diseases, and they are all nouns. A list of these words can be seen in Appendix C. In a medical context, nouns can, for example, describe a medical condition, a disease or a diagnosis. The system created in the current project only handles unigrams. If the system would have handled bigrams, adjectives describing the disease could be handled together with the disease, and thereby be relevant. However, in the current project, filtering out all non-nouns will yield a more interesting and accurate result since synonyms always belong to the same part-of-speech. The distribution of the different part-of-speech tags can be seen in Appendix B.

An excellent example of why the unigram model makes other parts-of-speech irrelevant is the word “kronisk (Eng: chronical),” which in the corpora often appear together with a disease, like “kronisk rinit (Eng: chronic rhinitis).” When trying to find the meaning of “rinit,” “kronisk” does not add information other than that it is a long-term disease. However, since they often appear together, the system might give them a high similarity, indicating that they have a similar meaning. The available corpora do not come with part-of-speech tags – requiring this to be done prior to further analysis. In general, there are many ways to perform the part-of-speech tagging, but the use of Swedish limits the possibilities. For this project, Stagger was used.

Stagger is a Swedish part-of-speech tagger. It is trained on the Stockholm-Umeå Corpus (SUC) and uses 22 different part-of-speech tags. SUC is a corpus where each of its one million tokens has manually been assigned a part-of-speech tag. The method that Stagger uses is called average perceptron (Östling, 2013).

The tag of interest in the current work is “NN,” which represents a noun. All the SUC tags and their meaning can be seen in Appendix B. If Stagger does not recognize a word as Swedish, it gets the tag “UO.” Since some of the medical terms might be rare, there is a possibility that
they, even though they are Swedish, get tagged as “UO.” Furthermore, upon inspecting the results, it was noticed that medical terms were tagged as “RG,” meaning it is a cardinal number. Therefore, the tags considered necessary to keep were “NN,” “UO” and “RG.” All other words were excluded from the analysis.

In some cases, Stagger gives the same word different part-of-speech tags, depending on its context. For this purpose, a function was created, that counted the number of times each word was tagged with a certain part-of-speech tag. Then, this function returns the part-of-speech tag that the word most frequently gets tagged with. An example of this is the word “faryngit” that is tagged as a noun 230 times and as a cardinal number 176 times. The returned part-of-speech tag would, in this case, be “NN,” which is correct.

The exclusion was done by adding the undesirable words to the stop word list that was mentioned in 3.3.2. However, it may be worth to know that this method of filtering out irrelevant words is not perfect in any way. For example, some of the more medical adjectives were not recognized by Stagger and could not be removed.

3.3.5 Stemming and lemmatization

In natural language processing, stemming is a good way to do the data more manageable. When dealing with large pieces of text, it is very likely that the same word will appear in different forms. That is, for example, with different inflectional forms. On the task of creating a co-occurrence matrix, words with the same word stem might appear as unique words. This redundancy can be avoided by, from the beginning, keeping only the word stems instead of the original words. Thus, reducing the number of unique words.

Before performing stemming or lemmatization, both “lungblåsor” (Eng: pulmonary alveoli) and “lungblåsorna” (Eng: the pulmonary alveoli) occurred in the same context as “lungemfysem” (Eng: pulmonary emphysema). Thus, these words’ relation to “lungemfysem” will be counted individually, although they are inflectional forms of the same word. Distinguishing between these words is unnecessary since it is only the meaning that is of interest in the analysis.

The are many different ways to perform the stemming. In the quanteda package, there is a built-in function for this. However, this function does not work in a way satisfactory to this project. The most straightforward algorithm for stemming is to remove suffixes of words, following a set of rules. For English, this could be replacing the suffix “ies” with “i” and “s” with nothing (Manning, Ragahvan, & Schutz, 2009). This set of rules will be different for different languages. It is evident that this method is not perfect, because not all inflections of words are that simple.

An example of when this does not work is for the Swedish word “bättre” (Eng: better). The word stem of “bättre” is “bra” (Eng: good), but just removing the suffix, which in this case is “e” will result in “bättr,” which is not a Swedish word. The following code, using the NLTK package for Python, shows the downside of a crude stemming algorithm.

```python
from nltk.stem import SnowballStemmer
stemmer = SnowballStemmer('Swedish')
stemmer.stem('bättre')
```

```
'bättr'
```
Another approach to this problem is using Stagger, that was mentioned in 3.3.4. In addition to tagging the words, Stagger provides the lemma for every word it recognizes. After tagging the text, it is possible to identify each word and its lemma, and then replace all words in the text with its corresponding lemma. Though, Stagger is not able to tag and find the lemma for all the words in the eCare corpus. In these cases, the original word is kept as it is. This problem occurs for some of the more obscure medical terms. The following code shows the better performance of a function utilizing the lemmas provided by Stagger.

```python
get_lemma('bättre')
```

```python
>>> 'bra'
```

Since Stagger has already been used for the part-of-speech tagging, lemmatizing the corpus is not more computationally costly than stemming. That is because all the lemmas have already been extracted and it is just to replace its original word in the corpus.

### 3.4 Finding semantically related words

As mentioned in section 3.3.2, the textstat_simil function provided several possibilities for calculating similarities. Thus, one method had to be decided upon, and this had to be done systematically. Therefore, using a window size of 10, each available similarity measurement was used to extract five words to every SNOMED term and was then compared to the others. The resulting data was intrinsically evaluated, to make visible the one method returning the most suitable words.

The similarity metrics compared were cosine, Jaccard, eJaccard, Dice, and eDice. For every synonym, the method received two points, and for every related word it received one point. All words were evaluated against 1177.se, to know whether they were synonyms or related. There were, in total, 100 words being evaluated for every method. This intrinsic evaluation resulted in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Cosine</th>
<th>Jaccard</th>
<th>eJaccard</th>
<th>Dice</th>
<th>eDice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58</td>
<td>79</td>
<td>67</td>
<td>80</td>
<td>68</td>
</tr>
</tbody>
</table>

This comparison in Table 2 elicits the difference between the methods, and it is clear that using Dice or Jaccard yields the best results. In all cases, tf-idf was used for weighting. That weighting method was chosen due to a drastically better performance than the other available weighting methods.

As a conclusion of this section, the following experimental setting was used for the final results:

- Window size: 10
- Similarity metric: Dice
- Weighting scheme: Tf-idf
- Number of output words: 5

When run, the system outputs one CSV file for every SNOMED term. These CSV files contain the five most similar words and their corresponding similarity. Using the words and values in these files, it was possible to perform the analysis.
In Figure 2, all the steps in extracting the candidate related terms can be seen.

![Figure 2. A description of the procedure of extracting the candidate related terms.](image)

### 3.4.1 Corpus expansion

During some initial testing, it was concluded that the corpus was too small. This could be observed by intrinsically evaluating the results. An example is the term “mastit” (Eng: *mastitis*) which should be closely related to “mjölkstockning” (Eng: *breastfeeding difficulties*). Although, the most similar word in the output of the system was “bröst.” “Mjölkstockning” was not even among the top ten most similar words. The reason for this turned out to be the corpus. The number of occurrences of the word “mjölkstockning” was two, whereas for “bröst” it was 42. Thus, explaining the skewed result.

For expanding the corpus, the software BootCat was used. BootCat can take in a list of words and create a search query for each word. Ten words from SNOMED was chosen for this list of terms. The search queries, containing the desired terms, are then used with Google. BootCat uses the first ten URLs from each Google result, downloads the text from them and then automatically creates a corpus. The corpus is structured as one text file for each website. The search was limited to only .se domains and sv.wikipedia.org, to avoid foreign and machine translated texts. Furthermore, the domain 1177.se was excluded from the search, to enable evaluation against this site at a later stage. When the corpus had been extracted and cleaned, it was concatenated with the original corpus, adding 174 new documents. Although, before concatenating it with the original corpus, it was thoroughly cleaned using regular expressions. This cleaning proved to be highly necessary due to the extent of which non-Unicode characters appeared in the text.

### 3.5 Evaluation

Although two words may have a high cosine similarity, it is not certain that they, in fact, are synonyms or even semantically related. To check whether two words are synonyms or related, one can either use a gold standard or use crowdsourcing. Evaluating a model against a gold standard is very common in the research field of natural language processing. The benefit of using a gold standard is that it is fast. It is comparable to finding a translation of a word in a glossary. For evaluation of Swedish medical synonyms, there is no already created gold standard. As a result, this approach requires a gold standard to be built prior to the analysis.

The crowdsourcing alternative comes with many benefits. Firstly, it is not dependent on an exact definition of synonymy. As in the case of Kann and Rosell (2005), the users could rely on their own understanding of synonymy. Secondly, crowdsourcing distributes the workload over a large crowd, minimizing the amount of work a single person has to put in. However, all
participants in the current work had to grade the same set of words due to the limited time and the magnitude of the available data.

When asking people to find words related to each other, one must put much thought into how the question is asked. There are several different ways to which words can be related. The words “car” and “automobile” are semantically related, but so are “car” and “bus” since they are hyponyms. A great way to do this is not telling them explicitly to identify synonyms. Instead, to widen their perspective, it is better instructing them to identify semantically related words and also to provide examples of the different kinds of relations.

As Kann and Rosell (2005) mentioned, there is a risk that the participants get irritated if the word pairs presented have no relation whatsoever. Ensuring that the system returns word pairs of good quality is of great importance for keeping the participants engaged throughout the questionnaire.

3.5.1 Questionnaire

In the light of the previously mentioned crowdsourcing approach, a questionnaire was created. The questionnaire was structured in a way similar to that of Kann and Rosell (2005). However, the rating scale had to be customized for the current study, using a scale from one to five and an additional “I don’t know” option to catch the cases where a participant does not know the meaning of a word. Every part started with a question in Swedish, asking the users, for example: “That what extent do the following words have a meaning similar to that of “anemi”?”. Below were the words to evaluate, each connected to the grading scale. Instructions were placed in the header and contained information about the definition of synonymy and how to interpret the grading scale.

The words in the questionnaire were the output of the system described in section 3.3. For each of the 15 SNOMED terms, five related words were considered reasonable for evaluation. There was also a need for controlling whether the users were guessing and answering without knowing the meaning of the words being evaluated. A participant who consistently answers that the control word has a similar meaning to the word being evaluated might not be reliable. Therefore, a randomly chosen word from the corpus was assigned to each set of five words. The random words were chosen from the total of all the extracted context windows from each SNOMED term. That is because they had to meet the same requirements as all the other words in the analysis, by being stemmed and nouns. From these extracted windows, there was a total of 59092 words and the script for extracting the words randomly also made sure that the same word never was chosen twice. Hence, the participants got to evaluate, in total 90 words, out of which 15 were control words.

Kann and Rosell (2005) used a scale where the ratings went from “I do not agree” to “I fully agree.” In the current work, however, it was decided that each rating should have an explicit definition, to make it easier for the participants. Thus, the definitions agreed upon was as follows:

1. The words have entirely different meanings (ex: nästäppa & ryggbesvär, Eng: nasal congestion & back problem)
2. The words can be related to each other (ex: nästäppa & symptom, Eng: nasal congestion & symptom)
3. The words are related and are often used together (ex: nästäppa & förkylning, Eng: nasal congestion & cold)

4. The words are strongly related and can sometimes replace each other (ex: nästäppa & snuva, Eng: nasal congestion & rhinitis)

5. The words are synonyms and often replaces each other (ex: nästäppa & nasalobstruktion, Eng: nasal congestion & nasal obstruction)

Although Kann and Rosell (2005) used a scale from zero to five, having only a scale from one to five was believed to be sufficient for the current work. Otherwise, the definitions would overlap each other, making it more difficult for the participants to answer. Both the questions and the alternatives were ordered randomly, using the built-in functionality in Google Forms, to avoid any order effect. However, the first question was always whether the participants were or had been working in medicine, or had studied medicine. The structure of the questionnaire can be seen in Appendix D.

Furthermore, Kann and Rosell (2005) encouraged the participants to answer intuitively – to base their answers on their intuitive idea of each word’s meaning. This approach would have been useful. Although, the purpose of the questionnaire was rather to evaluate existing results than to build a dictionary of synonyms. Additionally, Kann and Rosell (2005) based their results on 2.1 million ratings, whereas in the current work, there was not enough time to collect that amount of data. Encouraging the participants to answer intuitively could, therefore, give misleading results.

The questionnaire was distributed to the participants through an URL published in a Facebook group for nurses in Sweden and was accessible for seven days before it was closed. To make sure that the questionnaire would look the same for all participants, the questionnaire website was tested with the four most commonly used web browsers. Creating the questionnaire using Google Forms also made it responsive, providing the possibility to submit the answers through a smartphone.

Before the questionnaire was published, a pilot test was conducted. This test was essential to know whether the participants would understand the rating scale and what was asked of them. There was a suspicion that the participants would answer binary – that they would only answer either one or five. This binary answering would, in turn, lead to a less useful evaluation, since only synonyms and non-synonyms could be checked, not somewhat related words. After the pilot test, examples of word pairs were added to each alternative, making it easier to understand the rating scale. From the beginning, the plan was to have 20 SNOMED terms in the questionnaire, but the pilot test showed that this would take more than ten minutes. Considering this excessive time consumption, five terms were removed, resulting in a questionnaire of 15 terms. Furthermore, the pilot test revealed some concerns about the participants getting embarrassed by not knowing the meaning of the words. This resulted in adding a paragraph to the questionnaire description, telling the participants that they were not the subject of this study and that their answers would only be used to evaluate the implemented system. All participants were anonymous, and the only information that could be considered as personal was the yes-or-no question about their connection with the healthcare system.

3.5.2 Participants
The participants were chosen with respect to their putative medical knowledge. That is, people with occupations exposing them to medical terminology were considered suitable. The
SNOMED terms chosen for evaluation were considered too specialized for lay people to answer. The purpose of the questionnaire is not to find out whether people know the meaning of the terms. Instead, its purpose is to evaluate the implemented system. Thus, the selection of participants was focused on those that were believed to know the meaning of the terms. A control question was placed at the beginning of the questionnaire, to check whether the participants were or had been working in healthcare, or if they were studying or had studied a healthcare related subject. Since the questionnaire was published in a Facebook group for nurses in Sweden, it is reasonable to assume that most of the participants were nurses, although this was not further investigated. The total number of participants was 239, out of which 16 had to be excluded. Four participants had to be excluded due to answering no or not answering at all on the control question, and 12 participants due to insufficient answering (leaving more than 15 words without any answer).

3.5.3 Inter-rater reliability

Knowing the reliability of the raters (participants in the evaluation) is crucial for comparing their answers with the system output’s Dice similarity. For calculating this reliability, Fleiss Kappa was used. Fleiss kappa is used for measuring the agreement of several raters (Fleiss, 1971). The formula for calculating Fleiss kappa is shown in Equation 13, with its variables explained in Equations 9, 10, 11 and 12. Both the item agreement and category agreement might be useful to present in addition to the Fleiss kappa since they can provide a way of assessing how well the participants agreed on a certain item or category.

\[
ItemAgreement = \frac{\sum_{i=1}^{6} Category(i)^2 - NumberOfResponses}{NumberOfResponses \times (NumberOfResponses - 1)} \tag{9}
\]

\[
CategoryAgreement = \frac{TotalResponsesInCategory}{NumberOfItems \times NumberOfParticipants} \tag{10}
\]

\[
\bar{P} = Average \ item \ agreement \tag{11}
\]

\[
\bar{P}_e = \sum_{i=1}^{6} CategoryAgreement(i)^2 \tag{12}
\]

\[
\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{13}
\]

Landis and Koch (1977) propose a way of interpreting Kappa values. The proposal is as in Table 3. Worth to note is that this proposal contains a large portion of arbitrariness and in no way provides exact definitions. Despite this, it can be a useful tool when discussing the results.

<table>
<thead>
<tr>
<th>Kappa Statistic</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>Poor</td>
</tr>
<tr>
<td>0.00-0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81-1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

Table 3. Landis and Koch (1977) proposal of Kappa interpretation.
3.5.4 Correlation
For correlation, Kendall’s tau was used. This choice is motivated by the data not being normally distributed. Kendalls’s tau uses the difference between the number of concordant and discordant pairs. This is divided by the total number of pairs (Kendall, 1938).
4 Results

The results presented in this chapter are, firstly, the system output, that is the 15 SNOMED terms, their five most similar words, and their corresponding similarity. Secondly, the system output was a subject of human evaluation, and the results of this evaluation are the major part of this section since they provide a way of assessing the system’s performance.

4.1 System output

The system output for all the 15 SNOMED terms chosen for evaluation and their Dice similarity can be seen in Table 4, with the candidate related term in the left column and its corresponding Dice similarity in the right column. This table is what can be seen as thesaurus mentioned in section 2.7, although some of the words seem not to be suitable for this list. English translations of the words can be seen in Appendix E.

Table 4. The system output, showing each SNOMED term’s five most similar words as well as the Dice similarity.

<table>
<thead>
<tr>
<th>artrit</th>
<th>bronkit</th>
<th>depression</th>
<th>dermatit</th>
</tr>
</thead>
<tbody>
<tr>
<td>reumatoid</td>
<td>0.56</td>
<td>luftväg</td>
<td>0.35</td>
</tr>
<tr>
<td>psoriasisartrit</td>
<td>0.21</td>
<td>kol</td>
<td>0.28</td>
</tr>
<tr>
<td>knä</td>
<td>0.19</td>
<td>lunga</td>
<td>0.27</td>
</tr>
<tr>
<td>symptom</td>
<td>0.17</td>
<td>pneumoni</td>
<td>0.25</td>
</tr>
<tr>
<td>barn</td>
<td>0.17</td>
<td>luftvägsinfektion</td>
<td>0.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>emfysem</th>
<th>faryngit</th>
<th>hemicrania</th>
<th>hyperglykemi</th>
</tr>
</thead>
<tbody>
<tr>
<td>lunga</td>
<td>0.48</td>
<td>hals</td>
<td>0.50</td>
</tr>
<tr>
<td>lungsjukdom</td>
<td>0.34</td>
<td>slemhinna</td>
<td>0.27</td>
</tr>
<tr>
<td>obstruktiv</td>
<td>0.30</td>
<td>sjukdom</td>
<td>0.25</td>
</tr>
<tr>
<td>lungemfysem</td>
<td>0.29</td>
<td>svalg</td>
<td>0.25</td>
</tr>
<tr>
<td>lungvävnad</td>
<td>0.21</td>
<td>behandling</td>
<td>0.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hypotoni</th>
<th>pneumoni</th>
<th>prostatit</th>
<th>rinit</th>
</tr>
</thead>
<tbody>
<tr>
<td>blodtryck</td>
<td>0.61</td>
<td>crp</td>
<td>0.27</td>
</tr>
<tr>
<td>ortostatisk</td>
<td>0.38</td>
<td>bronkit</td>
<td>0.25</td>
</tr>
<tr>
<td>postural</td>
<td>0.28</td>
<td>kol</td>
<td>0.25</td>
</tr>
<tr>
<td>hjärtfrekvens</td>
<td>0.24</td>
<td>otit</td>
<td>0.18</td>
</tr>
<tr>
<td>blodvolym</td>
<td>0.21</td>
<td>pneumoniae</td>
<td>0.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>schizofreni</th>
<th>sinuit</th>
<th>tonsillit</th>
</tr>
</thead>
<tbody>
<tr>
<td>psykos</td>
<td>0.28</td>
<td>bihåleinflammation</td>
</tr>
<tr>
<td>beteende</td>
<td>0.20</td>
<td>otit</td>
</tr>
<tr>
<td>tonär</td>
<td>0.18</td>
<td>öli</td>
</tr>
<tr>
<td>hjärnvolym</td>
<td>0.18</td>
<td>polyp</td>
</tr>
<tr>
<td>mathalon</td>
<td>0.18</td>
<td>karies</td>
</tr>
</tbody>
</table>
Only five words for each target word was extracted, although more words could be related to the target word. This limitation was set to not overwhelm the questionnaire participants with words to rate. Hence 75 words and an additional 15 control words were enough for the questionnaire to take a maximum of ten minutes to answer.

The similarity calculated in R was, in general, rather low. As much as 91% of the words had a Dice similarity that was less than 0.50. The word with the highest similarity was the word pair “hemicrania-continua.” The lowest found similarity was 0.17, and three word pairs got a similarity this low.

There is a difference in how much text the corpus contains for each SNOMED term. When extracting the context window for each word, this difference becomes salient and is shown in Table 5.

<table>
<thead>
<tr>
<th>Word</th>
<th>Number of words in window</th>
</tr>
</thead>
<tbody>
<tr>
<td>artrit</td>
<td>5542</td>
</tr>
<tr>
<td>bronkit</td>
<td>5174</td>
</tr>
<tr>
<td>depression</td>
<td>11544</td>
</tr>
<tr>
<td>dermatit</td>
<td>7377</td>
</tr>
<tr>
<td>emfysem</td>
<td>3667</td>
</tr>
<tr>
<td>faringit</td>
<td>8270</td>
</tr>
<tr>
<td>hemicrania</td>
<td>1101</td>
</tr>
<tr>
<td>hyperglykemi</td>
<td>2180</td>
</tr>
<tr>
<td>hypotoni</td>
<td>758</td>
</tr>
<tr>
<td>pneumoni</td>
<td>2465</td>
</tr>
<tr>
<td>prostatit</td>
<td>3482</td>
</tr>
<tr>
<td>rinit</td>
<td>2551</td>
</tr>
<tr>
<td>schizofreni</td>
<td>1153</td>
</tr>
<tr>
<td>sinuit</td>
<td>1402</td>
</tr>
<tr>
<td>tonsillit</td>
<td>2426</td>
</tr>
</tbody>
</table>

### 4.2 Evaluation results

In the analysis, each rating is represented by a number, as follows:

1. The words have entirely different meanings
2. The words can be related to each other
3. The words are related and are often used together
4. The words are strongly related and can sometimes replace each other
5. The words are synonyms and often replaces each other

#### 4.2.1 Inter-rater agreement

The agreement among the participants, calculated using Fleiss kappa, was $\kappa = 0.28$. According to the proposed interpretation in Table 3, this is a fair agreement.
There was a difference of agreement among the categories, shown in Table 6, where the highest agreement was in the lowest rating category. The category with the lowest agreement was the one stating that the words were synonyms. These values were calculated using the formula in Equation 10.

**Table 6. The agreement rate of each category.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>The words have entirely different meanings</td>
<td>0.29</td>
</tr>
<tr>
<td>The words can be related to each other</td>
<td>0.24</td>
</tr>
<tr>
<td>The words are related and are often used together</td>
<td>0.25</td>
</tr>
<tr>
<td>The words are strongly related and can sometimes replace each other</td>
<td>0.08</td>
</tr>
<tr>
<td>The words are synonyms and often replaces each other</td>
<td>0.06</td>
</tr>
<tr>
<td>I don’t know</td>
<td>0.08</td>
</tr>
</tbody>
</table>

### 4.2.2 Most similar words

The most similar words to every SNOMED term, based on their average user rating, was extracted and is presented in Table 7 together with the corresponding Dice similarity.

**Table 7. The highest user rated candidate related term for every SNOMED term.**

<table>
<thead>
<tr>
<th>SNOMED term</th>
<th>Candidate term related</th>
<th>Average user rating</th>
<th>Dice similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>artrit</td>
<td>psoriasisartrit</td>
<td>3.28</td>
<td>0.21</td>
</tr>
<tr>
<td>bronkit</td>
<td>luftvägsinfektion</td>
<td>3.84</td>
<td>0.24</td>
</tr>
<tr>
<td>depression</td>
<td>ångest</td>
<td>2.94</td>
<td>0.18</td>
</tr>
<tr>
<td>dermatit</td>
<td>eksem</td>
<td>3.51</td>
<td>0.43</td>
</tr>
<tr>
<td>emfysem</td>
<td>lungemfysem</td>
<td>4.35</td>
<td>0.29</td>
</tr>
<tr>
<td>faryngit</td>
<td>svalg</td>
<td>2.93</td>
<td>0.25</td>
</tr>
<tr>
<td>hemicrania</td>
<td>huvudvärk</td>
<td>3.26</td>
<td>0.59</td>
</tr>
<tr>
<td>hyperglykemi</td>
<td>diabetes</td>
<td>3.22</td>
<td>0.37</td>
</tr>
<tr>
<td>hypotoni</td>
<td>blodtryck</td>
<td>3.32</td>
<td>0.61</td>
</tr>
<tr>
<td>pneumoni</td>
<td>pneumoniae</td>
<td>3.95</td>
<td>0.18</td>
</tr>
<tr>
<td>prostatit</td>
<td>prostatakörtel</td>
<td>2.85</td>
<td>0.29</td>
</tr>
<tr>
<td>rinit</td>
<td>nästätta</td>
<td>3.47</td>
<td>0.29</td>
</tr>
<tr>
<td>schizofreni</td>
<td>psykos</td>
<td>3.09</td>
<td>0.28</td>
</tr>
<tr>
<td>sinuit</td>
<td>bihåleinflammation</td>
<td>4.72</td>
<td>0.29</td>
</tr>
<tr>
<td>tonsillit</td>
<td>halsfluss</td>
<td>4.39</td>
<td>0.38</td>
</tr>
</tbody>
</table>

### 4.2.3 System performance

The system output was divided into intervals with steps of size 0.05. Figure 3 shows the results and makes visible which interval that provides the highest rated word pairs. In this case, a Dice similarity between 0.35 and 0.39 had the highest average rating. There was no value below 0.17 and only two values between 0.60 and 1.00. Thus, the last interval has a step of size 0.40. The error bars show the standard deviation, where the highest standard deviation is in the interval 0.15-0.19, whereas the lowest is in the interval 0.50-0.54.
Figure 3. The performance of the system at specific intervals.

In Table 8, it is possible to observe the difference of the number of values within each interval.

Table 8. The values used in Figure 3.

<table>
<thead>
<tr>
<th>Relatedness interval</th>
<th>Human rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>0.15-0.19</td>
<td>1.99</td>
</tr>
<tr>
<td>0.20-0.24</td>
<td>2.32</td>
</tr>
<tr>
<td>0.25-0.29</td>
<td>2.65</td>
</tr>
<tr>
<td>0.30-0.34</td>
<td>2.60</td>
</tr>
<tr>
<td>0.35-0.39</td>
<td>3.23</td>
</tr>
<tr>
<td>0.40-0.44</td>
<td>2.68</td>
</tr>
<tr>
<td>0.45-0.49</td>
<td>1.91</td>
</tr>
<tr>
<td>0.50-0.54</td>
<td>2.76</td>
</tr>
<tr>
<td>0.55-0.59</td>
<td>2.76</td>
</tr>
<tr>
<td>0.60-1.00</td>
<td>2.80</td>
</tr>
</tbody>
</table>

The average rating of all the words, with the control words excluded, was 2.48 (SD = 0.80).

4.2.4 Control words

Out of the 15 control words, the average rating was 1.19 (SD = 0.23). The purpose of the control words was to have a way of controlling whether the participants was just guessing or if they actually knew the meaning of the target words. From the value 1.19, it is possible to conclude that the participants had some knowledge of the SNOMED terms, and that they therefore knew that the control words had no relation to the target words. This makes their answers more reliable.

4.2.5 “I don’t know” answers

Table 9 shows the ten word pairs with the highest number of participants answering “I don’t know.” Out of these word pairs, five belong to the SNOMED term “hemicrania.”
Table 9. Percentage of “I don’t know” answers.

<table>
<thead>
<tr>
<th>SNOMED term</th>
<th>Candidate related term</th>
<th>Percentage of “I don’t know” answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>hemicrania</td>
<td>indomethacin</td>
<td>72%</td>
</tr>
<tr>
<td>hemicrania</td>
<td>continua</td>
<td>68%</td>
</tr>
<tr>
<td>schizofreni</td>
<td>mathalon</td>
<td>61%</td>
</tr>
<tr>
<td>hemicrania</td>
<td>paroxysmal</td>
<td>52%</td>
</tr>
<tr>
<td>hemicrania</td>
<td>info</td>
<td>45%</td>
</tr>
<tr>
<td>hemicrania</td>
<td>huvudvärk</td>
<td>40%</td>
</tr>
<tr>
<td>hypotoni</td>
<td>postural</td>
<td>33%</td>
</tr>
<tr>
<td>dermatit</td>
<td>perioral</td>
<td>14%</td>
</tr>
<tr>
<td>sinuit</td>
<td>öli</td>
<td>9%</td>
</tr>
<tr>
<td>sinuit</td>
<td>polyp</td>
<td>8%</td>
</tr>
</tbody>
</table>

4.2.6 Kendall’s tau

There was a significant relationship between the average rating of the words and their Dice similarity. A Kendall’s tau coefficient test showed the following: $r_T = .28$, $p < .001$. 

5 Discussion

Although the system provided candidate related terms of which the Dice similarity was significantly correlated with the human ratings, much could be done to improve the system. Therefore, it is of great importance to identify what factors may have had a negative impact on the results. The following chapter will discuss the results, what may have affected it, and suggestions for improving the system, and further research.

5.1 Main results

The correlation between the average human rating and the average Dice similarity was, as previously mentioned, \( r = .28, p < .001 \). While the correlation is significant, it is rather weak. What this correlation means is that the higher Dice similarity a word gets, the higher average human rating it gets. However, much can be said about these results. Firstly, the Dice similarity was in general rather low. As mentioned in section 4.1, 91% of all word pairs got a Dice similarity less than 0.5. Words that are synonyms, or at least by the participants considered closely related should have a Dice similarity closer to 1 than 0. The overall quite low Dice similarity may be a reason for the weak correlation.

A further example of the low Dice similarity is for the word pair “emfysem-lungemfysem” (Eng: emphysema-pulmonary emphysema), where the average user rating is 4.35, whereas the Dice similarity was only 0.29. What this rating means it that the participants perceived the similarity as somewhere in between The words are strongly related and can sometimes replace each other, and The words are synonyms and often replaces each other. The word “lungemfysem” is a compound of the two words; “lunga” and “emfysem.” Thus “lungemfysem” is a certain kind of emphysema, which makes it clear that they have a strong semantic relationship. It is difficult to decide on whether they are synonyms or just strongly related, but what is clear is that the participants were more accurate on rating these words than the system was.

5.2 Limited corpus

The total amount of words used as input was 789043. Table 7 shows the highest rated candidate related term for every SNOMED term. Among these, the word pair which was least similar according to the participants, with an average rating of 2.85, was “prostatit-prostatakörtel” (Eng: prostatitis-prostate gland). The word “prostatit” means inflammation of the prostate gland and has the synonym “blåshalskörtelinflammation.” In a best-case scenario, the system would have found this synonym. However, the word “blåshalskörtelinflammation” only occurs twice in the context window extracted around “prostatit.” The word “prostatakörtel” on the other hand occurs 116 times. A comparison with the highest rated word pair, “sinuit-bihåleinflammation” (Eng: sinusitis-sinus infection) emphasizes how the word occurrence affects the similarity. “Bihåleinflammation” occurs 68 times together with “sinuit.” A reason for “blåshalskörtelinflammation” rarely occurring might be that it is not, in general, being used that often. It might be that “prostatit” more often is described with an expression containing several words, and by limiting the analysis to unigrams, this definition is impossible to catch.

Section 4.1, Table 5 shows that there was a considerable size difference of the context windows, where the window for “depression” was more than 15 times the size of that of “hypotoni.” The

---

4 According to the definition at: https://mesh.kib.ki.se/term/D011472/prostatitis
fewer words in a context window, the more impact each word has on the similarity measure. Preferably, there would have been a more even distribution of texts among all the SNOMED terms.

5.3 Performance intervals
As can be seen in Figure 3, a higher relatedness score does not necessarily mean that the human rating will be higher. However, it is crucial to know that the number of words in each interval differ radically, as Table 8 shows. For example, the reason for the interval 0.45-0.49 exhibiting such low human rating is two word pairs (“hemicrania-info” and “emfysem-lunga”) getting the ratings 1.14 respectively 2.68. Since this interval only contains two word pairs, it is not possible to say that the system, in general, performs poorly in this interval. For concluding this, more data would be needed.

Ideally, there would be a linear increase of human rating visible in Figure 3. However, there was a significant correlation between the average human rating and average Dice similarity when calculating Kendall’s tau coefficient. The reason for the graphical representation in Figure 3 was to find a threshold for when the relatedness score would provide good enough word pairs, but this proved to be difficult. According to Figure 3, the performance is not much better when the Dice similarity is between 0.60 and 0.90 than when it is between 0.25 and 0.29. One reason for this could be the lack of data. The pilot test mentioned in section 3.5.1 resulted in a decrease from 20 to only 15 SNOMED terms due to excessive time consumption. Testing only 15 SNOMED terms resulted in the participants only rating 75 terms, with the control words excluded. It is not possible to generalize the results with only 75 terms. As can be seen in Table 8, if the average rating of an interval is based on only two terms, it is not reliable. Especially not in comparison with another interval containing 21 values. Furthermore, the error bars in Figure 3 show that the standard deviation is rather high in relation to the mean values for all intervals except 0.50-0.54. Although it is possible to observe a non-linear relation between the intervals, the mean values are not that reliable.

5.4 Unigram limitation
Limiting the system only to handle unigrams may have decreased the overall performance of the system. Some of the terms might have to be described by more than one word. An example of this is “hyperglykemi” (Eng: hyperglycemia) which does not have a synonym. Instead, it is best described as “onormalt hög blodsockernivå” (Eng: abnormally high blood glucose level).\(^5\) This is a trigram and would therefore not be found by the system in this work. In the case of “onormalt hög blodsockernivå,” the words “onormalt” and “hög” would have been removed in the noise removal stage, which is reasonable since they, taken out of context, have no apparent relation to “hyperglykemi.” The fact that there is no synonym to “hyperglykemi” explains why none of the candidate related terms got a higher human rating than 3.22. There are other medical terms to which there are no synonyms, only correctly described by multiple-word expressions. This lack of synonyms emphasizes the need for a model that can manage n-grams larger than unigrams.

As mentioned in the theory chapter, Smith et al. (2014) concluded that abbreviations frequently occur in medical texts. Among the candidate related terms used in the evaluation, there are at least two abbreviations; “öli” and “crp.” The two abbreviations both have bigram expansions;

\(^5\) According to the definition at: https://mesh.kib.ki.se/term/D006943/hyperglycemia
“övre luftvägsinfektion” and “C-reaktivt protein.” When taken out of context, these abbreviations might be hard to know the meaning of. A workaround for this problem could be to build a system that can expand abbreviations, and then can find relations between unigrams and bi-grams.

5.5 Noise

Although much effort was put into removing noise, the analysis revealed that not all noise was removed. As was stated in the method chapter, it was decided only to keep the nouns. However, some of the more complicated medical adjectives were left in the corpus. An example of this is the word “atopisk” (Eng: atopic) that often occur together with “dermatit” (Eng: dermatitis) as the expression “atopisk dermatit.” Stagger did not manage to tag “atopisk” correctly. All 88 occurrences of this word were tagged with “NN,” making it mistakenly included in the analysis. This inaccurate tagging was also the case with a few other words. As mentioned in section 3.3.4, synonyms always belong to the same part of speech. If an adjective is included in the results, one can be certain that this is not a synonym. Had these words been excluded from the analysis, it is entirely possible that the results would have been more accurate.

5.6 Evaluation method

The level of agreement among the questionnaire participants was 0.28, as mentioned in section 4.2.1. That number represents a fair agreement. What this means is that the ratings are somewhat unreliable. The reason for this rather low agreement could be the design of the questionnaire. As Table 6 shows, the agreement was more moderate for the higher rating categories. In plain text, this would mean that it was easier for the participants to agree on that words were non-similar, than when they were similar. Asking humans to put a rating on something always entails some arbitrariness. It is not certain that they know the meaning of the words, nor that they perceive the similarity rating in the same way. For each rating in the rating scale, an example was provided. These examples, however, was evidently not enough for the participants to be sure of what to answer. The highest, on average, rated word pair was “sinuit-bihåleinflammation” with an average rating of 4.72 (SD = 0.66). Although this is a rather clear case of synonymy, 49 participants did not rate them as synonyms. This mean that as much as a fifth of the participants, either did not know the meaning of the word or how to interpret the questionnaire.

The participants’ knowledge in Swedish was not investigated. Since a majority of the participants are nurses in Sweden, it is possible to assume that their knowledge in Swedish is sufficient for performing their work. However, it would have been good to include a question about the participants’ native language.

Table 9 shows that some words tended to get a lot of “I don’t know” answers. Among the top ten, word pairs with “hemicrania” occur five times. The fact that “hemicrania” occurs so many times makes it reasonable to conclude that the participants, in general, had trouble with understanding the meaning of this word. What is problematic about this is that it makes the evaluation less reliable. During the pilot test, measures were taken to ensure that the 15 SNOMED terms were commonly known by healthcare professionals. However, not enough actions were taken to avoid this knowledge problem. What could have been done to circumvent this problem is, firstly, to include more SNOMED terms in the evaluation, reducing the effect one single SNOMED term has on the results. Secondly, more healthcare professionals could have been included in the pilot study to ensure that the SNOMED terms were commonly known.
As mentioned in the theory chapter, there is a possibility of the participants getting irritated when presented with bad word pairs. Irritated participants might not want to finish the questionnaire or answer randomly, without thinking. Whether the participants got irritated in the current evaluation is difficult to investigate further. However, the average control word rating was 1.19, indicating that the participants were not answering by guessing. Moreover, only 12 participants had to be excluded due to insufficient answering. Most of the other participants answered almost every question.

Usually, systems like these are evaluated using a gold standard, resulting in measurements of accuracy, precision, and recall like the ones that Henriksson et al. (2012) presented. For calculating these measures, each candidate related term would have to be considered either right or wrong. Ideally, there would already exist a synonym list or thesaurus, and thus a candidate related term would be right if it existed in the corresponding SNOMED term’s thesaurus entry. The fact that no such thesaurus existed was the reason for using crowdsourcing instead.

5.7 Lay and expert difference
From the beginning, the corpus was split into one lay corpus and one specialized corpus. However, there was an immense difference of sizes between them, where the specialized was more than twice the size of the lay corpus. Furthermore, the corpus needed to be expanded with texts concerning some of the SNOMED terms that were lacking text. Since the added texts were not annotated, it was not possible to distinguish between the two categories upon the expansion. Based on this expansion, the original plan was abandoned, and thus both corpora were used as one.

5.8 Real-world usage
When replacing synonyms in texts, it is imperative that no information gets lost, especially in a medical context where correct information could be the difference between life and death. If a healthcare professional makes an entry in a patient’s journal, the patient still must receive the correct information, although a system has changed some of the words. As Keskisärkkä & Jönsson (2012) concluded, the choice of synonyms is heavily context-dependent, thus replacing synonyms must be done carefully and with much regard to the context. On the task of replacing synonyms, the performance of the current system would have to be improved and be a subject of extensive testing before using it on real-world tasks. Furthermore, to assess the performance on real-world tasks, access to actual medical records would be needed. This access, in turn, comes with confidentiality issues. Dealing with this sensitive and personal information is difficult and entails many ethical questions that have to be considered.

5.9 Further research
There is much room for improvement of the current work. Therefore, suggestions for further research, such as expanding the corpus and change of evaluation methods are discussed in the following section.

5.9.1 Gold standard evaluation
Creating a gold standard for evaluating the system might provide a more robust evaluation than when using crowdsourcing. This gold standard could be created using the Swedish version of MeSH. However, a problem with MeSH is that it only explicitly provides the synonyms for the medical terms. If, as in this work, other semantically related words also are of interest, one
would have to use the term descriptions in MeSH. Using these descriptions, preferably, a thesaurus could be created.

5.9.2 Corpus expansion
In the case of distributional semantics, more data leads to a more reliable result. For this reason, redoing this work with a bigger corpus would probably provide more accurate synonyms. Furthermore, the corpus expansion would provide the possibility of utilizing the functions in R for handling sparse matrices. In the current work, applying RI and LSA was not possible due to lack of data. It is easy to propose an expansion of the corpus, but the question is from where the data will come. Using Google to find Swedish websites containing the term “hypotoni” there are approximately 53000 hits. This number may sound high, but a significant portion of these websites contain machine-translated text and thus must be considered somewhat syntactically and semantically unreliable. Therefore, the great challenge of collecting a large enough Swedish corpus remains.

Furthermore, a good idea for further work is to expand the lay version of the corpus. Making the lay corpus equal to the specialized version could make a comparison of the two possible. It would be interesting to see what different words the system would find in the two different corpora.

5.9.3 More advanced methods
The results are good, considering the rather simple method used for extracting the related words. However, other, more advanced methods would probably yield a better result. Working with bigrams and trigrams would, for example, make the inclusion of adjectives relevant. Moreover, it would be possible to find accurate descriptions to the medical terms lacking a synonym. In the current work, the weighting scheme used was tf-idf because PPMI was not available in the quanteda package. Although, it would be interesting to see how the results would have changed when using PPMI.

As mentioned in section 3.3.2, handling a full co-occurrence matrix was too computationally heavy, and applying methods for dimensionality reduction was not possible. Instead, co-occurrence matrices were created from extracted context windows. For further research, it would be interesting to use RI or LSA and create a full co-occurrence matrix. That is the method used by Henriksson et al. (2012), and it proved to be successful.

5.9.4 Measuring comprehensiveness
Although the model used in this work managed to find some synonyms to medical terms, there was no measurement of whether the synonyms were easier to understand than the original terms. For further research, it would be a good idea to conduct an additional evaluation of the comprehensiveness of the extracted synonyms. It could be that the extracted synonym was just as difficult to understand as the original term, and in that case, replacing the word in a text would not improve the readability.

5.9.5 English corpus
The current corpus was constructed of Swedish medical texts. However, the system should be easy to customize for other languages such as English. Worth to note though is that word compounds are more commonly used in Swedish than in English. While “lungvävnad” is a unigram, its English equivalent is “lung tissue” and thus a bigram. This is the case when translating many of the Swedish words in the corpus. Had the model been implemented in English, it is possible
that the performance of a unigram model would decrease. Implementing the system in English could, however, increase the usefulness since English a more widely used language than Swedish. Furthermore, the original English version of MeSH is much bigger than its Swedish equivalent, thus making it easier to create an extensive gold standard.
6 Conclusion

The purpose of this work was to extract synonyms and related words from a medical corpus. It was indeed possible to extract related words from the given E-care corpus. The evaluation showed that although the system was far from optimal, it was able to extract words that are related. Since the average human rating of the extracted words was 2.48, that is somewhere in between The words can be related to each other and The words are related and are often used together. However, many measures could be taken to increase the performance of the system. One crucial part is not to limit the system to only unigrams since some medical terms need at least bigrams to be accurately represented. Moreover, a bigger and more evenly distributed corpus would provide more reliable results.

On the task of creating a thesaurus, the system was successful. The higher goal of the E-care@home project that this work has contributed to is to replace complicated medical terms with their lay synonyms. For solving this, there is a positive outlook, since a rather simple system, created in quite a short time, such as the one in this work, still manages to extract related words. Improving the performance is possible, and with little work on the method and a larger corpus, the system will likely provide good synonyms. When the system provides good enough synonyms, what remains is to assess the comprehensiveness of the extracted synonyms, making sure the synonym replacement makes the texts easier to understand. It will be interesting to follow the development in this area and what possibilities distributional semantics will provide. What is sure is that the future for this research area looks bright.
References


Appendix A

An example of the XML file structure, with the relevant information.

```xml
<text>
  <textHeader>
    <textLangUsage>
      <subLanguages>
        <sublang lay-annotator lay></sublang>
        <comment lay-annotator>[...]</comment>
        <sublang expert-annotator lay></sublang>
        <comment expert-annotator>[...]</comment>
      </subLanguages>
    </textLangUsage>
    <savedText>
      "Men det är inte som på film"
      Miriam Jaakola, 38, var 26 år när hon fick sin diagnos. Foto: Tomas Bergman
      Hon varken svår eller säger fula ord – men hon drabbas ofta av tics och har svår att tygla sina impulsor.
      Miriam Jaakola, 38, har Tourettes syndrom.
    </savedText>
  </textHeader>
</text>
```
Appendix B

The meaning of the part-of-speech tags and how frequent they occur in the corpus.

<table>
<thead>
<tr>
<th>POS tag</th>
<th>Meaning</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Noun</td>
<td>248955</td>
</tr>
<tr>
<td>VB</td>
<td>Verb</td>
<td>114542</td>
</tr>
<tr>
<td>PP</td>
<td>Preposition</td>
<td>95953</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>68067</td>
</tr>
<tr>
<td>AB</td>
<td>Adverb</td>
<td>51215</td>
</tr>
<tr>
<td>KN</td>
<td>Conjunction</td>
<td>37610</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>27920</td>
</tr>
<tr>
<td>PN</td>
<td>Pronoun</td>
<td>27142</td>
</tr>
<tr>
<td>RG</td>
<td>Cardinal number</td>
<td>24701</td>
</tr>
<tr>
<td>UO</td>
<td>Foreign word</td>
<td>19666</td>
</tr>
<tr>
<td>PC</td>
<td>Participle</td>
<td>17117</td>
</tr>
<tr>
<td>HP</td>
<td>Relative pronoun</td>
<td>12287</td>
</tr>
<tr>
<td>PM</td>
<td>Proper noun</td>
<td>11817</td>
</tr>
<tr>
<td>SN</td>
<td>Subordinating conjunction</td>
<td>10173</td>
</tr>
<tr>
<td>IE</td>
<td>Infinitive marker</td>
<td>8059</td>
</tr>
<tr>
<td>PL</td>
<td>Verb particle</td>
<td>4709</td>
</tr>
<tr>
<td>HA</td>
<td>Relative adverb</td>
<td>3935</td>
</tr>
<tr>
<td>PS</td>
<td>Possessive pronoun</td>
<td>2791</td>
</tr>
<tr>
<td>RO</td>
<td>Ordinal number</td>
<td>1018</td>
</tr>
<tr>
<td>HD</td>
<td>Relative determiner</td>
<td>318</td>
</tr>
<tr>
<td>IN</td>
<td>Interjection</td>
<td>207</td>
</tr>
<tr>
<td>HS</td>
<td>Possessive relative pronoun</td>
<td>58</td>
</tr>
</tbody>
</table>
## Appendix C

The 15 SNOMED terms chosen for analysis and their randomly assigned control words.

<table>
<thead>
<tr>
<th>SNOMED term</th>
<th>Control word</th>
</tr>
</thead>
<tbody>
<tr>
<td>artrit</td>
<td>förlust</td>
</tr>
<tr>
<td>bronkit</td>
<td>synnerhet</td>
</tr>
<tr>
<td>depression</td>
<td>form</td>
</tr>
<tr>
<td>dermatit</td>
<td>roflumilast</td>
</tr>
<tr>
<td>emfysm</td>
<td>tid</td>
</tr>
<tr>
<td>faryngit</td>
<td>sätt</td>
</tr>
<tr>
<td>hemicrania</td>
<td>år</td>
</tr>
<tr>
<td>hyperglykemi</td>
<td>tolkning</td>
</tr>
<tr>
<td>hypotoni</td>
<td>deltahepatit</td>
</tr>
<tr>
<td>pneumoni</td>
<td>symtom</td>
</tr>
<tr>
<td>prostatit</td>
<td>tand</td>
</tr>
<tr>
<td>rinit</td>
<td>röst</td>
</tr>
<tr>
<td>schizofreni</td>
<td>administration</td>
</tr>
<tr>
<td>sinuít</td>
<td>validerade</td>
</tr>
<tr>
<td>tonsillit</td>
<td>postpartumdepression</td>
</tr>
</tbody>
</table>
Appendix D

The structure of the questionnaire.

The first page shows a description of questionnaire and for what purpose the answers will be used.

**Medicinska synonymer och språkligt relaterade ord**


Resultatet från den här enkäten kommer att användas för att utvärdera ett automatiskt system för identifiering av medicinska synonymer och relaterade ord. Det är alltså inte du som svarar som studeras, utan dina svar kommer enbart att användas för att ha något att jämföra systemet med, då det i nuläget inte finns något facit.

Svara på en femgradig skala med följande alternativ:
1 - orden har helt olika betydelser (ex: nästäppa & ryggbesvär)
2 - orden kan vara relaterade till varandra (ex: nästäppa & symptom)
3 - orden är relaterade och används ofta tillsammans (ex: nästäppa & förkyning)
4 - orden är starkt relaterade och kan ibland ersätta varandra (ex: nästäppa & snuva)
5 - orden är synonymer och ersätter ofta varandra (ex: nästäppa & nasalobstruktion)

Om du inte vet ordens betydelse kan du svara "vet ej".

Enkäten tar ca 5-10 min att genomföra. Vid eventuella frågor, kontakta: gusca715@student.liu.se

Arbetar du eller har du arbetat inom vården alt. studerar/studerat något vårdrelaterat ämne?

- Ja
- Nej
Till vilken grad har följande ord en språklig betydelse som liknar "emfysem"?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunga</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Lungsjukdom</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Obstruktiv</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Lungomfysem</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Lungvävnad</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Tid</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
## Appendix E

The 90 Swedish words used in the questionnaire and their English equivalents.

<table>
<thead>
<tr>
<th>Swedish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>administration</td>
<td>administration</td>
</tr>
<tr>
<td>allergi</td>
<td>allergy</td>
</tr>
<tr>
<td>artrit</td>
<td>arthritis</td>
</tr>
<tr>
<td>atopisk</td>
<td>atopic</td>
</tr>
<tr>
<td>bakterie</td>
<td>bacterium</td>
</tr>
<tr>
<td>barn</td>
<td>child</td>
</tr>
<tr>
<td>behandling</td>
<td>treatment</td>
</tr>
<tr>
<td>besvär</td>
<td>encumbrance</td>
</tr>
<tr>
<td>beteende</td>
<td>behavior</td>
</tr>
<tr>
<td>bikhåleinflammation</td>
<td>sinusitis</td>
</tr>
<tr>
<td>blodtryck</td>
<td>blood pressure</td>
</tr>
<tr>
<td>blodvolym</td>
<td>blood volume</td>
</tr>
<tr>
<td>bronkit</td>
<td>bronchitis</td>
</tr>
<tr>
<td>bäckenbotten</td>
<td>perineum</td>
</tr>
<tr>
<td>continua</td>
<td>continua</td>
</tr>
<tr>
<td>CRP (C-reaktivt protein)</td>
<td>CRP (C-reactive protein)</td>
</tr>
<tr>
<td>deltahepatit</td>
<td>hepatitis delta</td>
</tr>
<tr>
<td>depression</td>
<td>depression</td>
</tr>
<tr>
<td>dermatit</td>
<td>dermatitis</td>
</tr>
<tr>
<td>diabetes</td>
<td>diabetes</td>
</tr>
<tr>
<td>eksem</td>
<td>eczema</td>
</tr>
<tr>
<td>emfysem</td>
<td>emphysema</td>
</tr>
<tr>
<td>episod</td>
<td>episode</td>
</tr>
<tr>
<td>faryngit</td>
<td>pharyngitis</td>
</tr>
<tr>
<td>form</td>
<td>form</td>
</tr>
<tr>
<td>förlust</td>
<td>loss</td>
</tr>
<tr>
<td>gas</td>
<td>gas</td>
</tr>
<tr>
<td>glukos</td>
<td>glucose</td>
</tr>
<tr>
<td>hals</td>
<td>throat</td>
</tr>
<tr>
<td>halsfluss</td>
<td>tonsillitis</td>
</tr>
<tr>
<td>halsont</td>
<td>sore throat</td>
</tr>
<tr>
<td>hemicrania</td>
<td>hemicrania</td>
</tr>
<tr>
<td>hjärnvolym</td>
<td>brain volume</td>
</tr>
<tr>
<td>hjärtfrekvens</td>
<td>heart rate</td>
</tr>
<tr>
<td>hud</td>
<td>skin</td>
</tr>
<tr>
<td>huvudvärk</td>
<td>headache</td>
</tr>
<tr>
<td>hyperglykemi</td>
<td>hyperglycemia</td>
</tr>
<tr>
<td>hypotoni</td>
<td>hypotension</td>
</tr>
<tr>
<td>indomethacin</td>
<td>indomethacin</td>
</tr>
<tr>
<td>info</td>
<td>info</td>
</tr>
<tr>
<td>insulin</td>
<td>insulin</td>
</tr>
<tr>
<td>karrie</td>
<td>caries</td>
</tr>
<tr>
<td>knä</td>
<td>knee</td>
</tr>
</tbody>
</table>
KOL (kroniskt obstruktiv lungsjukdom)
liv
luftväg
luftvägsinfektion
lunga
lungemfysem
lungsjukdom
lungvävnad
mathalon
mmol
måltid
näsa
nässlemhinnan
nästäppa
obstruktiv
ortostatisk
otit
paroxysmal
perioral
person
pneumoni
pneumoniae
polyp
postpartumdepression
postural
prostata
prostatakörtel
prostatit
psoriasisartrit
psykos
reumatoid
rinit
roflumilast
röst
schizofreni
sinuit
sjukdom
slemhinna
svalg
symtom
synnerhet
sätt
tand
tid
tolkning
tonsill
tonsillit
tonår

COPD (chronic obstructive pulmonary disease)
life
respiratory tract
respiratory tract infection
lung
pulmonary emphysema
lung disease
lung tissue
mathalon
mmol
meal
nose
nasal mucosa
nasal congestion
obstructive
orthostatic
otitis
paroxysmal
perioral
person
pneumonia
pneumonia
polyp
postpartum depression
postural
prostate
prostate gland
prostatitis
psoriatic arthritis
psychosis
rheumatoid
rhinitis
roflumilast
voice
schizophrenia
sinusitis
disease
mucous membrane
pharynx
symptom
particular
manner
tooth
time
interpretation
tonsil
tonsillitis
tens
<table>
<thead>
<tr>
<th>utslag</th>
<th>rash</th>
</tr>
</thead>
<tbody>
<tr>
<td>validerade</td>
<td>validated</td>
</tr>
<tr>
<td>ångest</td>
<td>anxiety</td>
</tr>
<tr>
<td>år</td>
<td>year</td>
</tr>
<tr>
<td>ÖLI (övre luftvägsinfektion)</td>
<td>URTI (upper respiratory tract infection)</td>
</tr>
</tbody>
</table>