Compound splitting of Swedish medical words

An evaluation of the *Compound Splitter* software

Ludvig Falck
8410140316

Omid Samadi
8809144754

Group 25

Department of Computer and Systems Sciences

Report for FMVEK/IC2003

Stockholm University

Summary

The focus of this research paper is Swedish compound words, and more specifically compound words related to medicine and health. With a set of words from this domain, our goal was to evaluate the performance of a general Swedish compound splitting software named *Compound Splitter*. We deemed the results of this research interesting for individuals that conduct indexation of medical documents. The output of *Compound Splitter* was compared to compound splits performed manually according to a set of rules. Our experiments show that *Compound Splitter* has an accuracy rate of 80.5%, a precision rate of 79.8% and a recall rate of 77.17% on this particular set of words. With these numbers in mind, we conclude that the software does have problems in dealing with certain words, specifically words unique to the world of medicine. All things considered however, we find that the performance is reasonably good as the software is meant to be used in conjunction with general Swedish compounds. The software could, perhaps with certain modifications, be used for splitting specialized words of this kind.

Keywords

Swedish compound splitting, medical words, *Compound Splitter, SWETWOL*

Contribution

The assignment instructions ask for percentage of contribution. Both authors have contributed roughly 50% each to all the chapters in the report.

Table of Contents

1 Introduction

1.1 Background

Compounding words, which is the joining of more than one word to form a new one, is rather common in the Swedish language. We encounter these words every day and use them frequently. "Arbetsmiljö" (work environment) is an example of a Swedish compound which is made up of the two individually meaningful words "arbete" (work) and "miljö" (environment). Correctly splitting a compound word is a task that is carried out for several different reasons. In certain cases it may improve one's understanding of the compound by simply interpreting the meaning of each component. In computational linguistics, the task of compound splitting is also a common one (Koehn & Knight, 2003). Here the aim is among other things to improve the quality of search engines and machine translation systems. When building search engines, the task of correctly splitting compounds is an essential strategy where the compound as a whole, along with its individual components must produce relevant search results (Ahlgren & Kekäläinen, 2007).

The focus of our research lies not with general Swedish compound words, but rather those frequently used in the medical world. An example of such a compound word is "halsinfektion" (throat infection), with the components "hals" (throat) and "infektion" (infection). Compound splitting is of particular interest for the medical area because it allows one to enhance indexation of extensive medical documents (Dujols, Aubas, Baylon, & Grémy, 1991).

1.2 Research area

Compound splitting is not always as simple as with the previous two examples. This is particularly true when analyzing words from medical documents. Manually splitting the compound word "koronarangiografi" is not necessarily straight forward. Fortunately, tools such as *SWETWOL* (Karlsson, 1992) and *Compound Splitter* (Sjöbergh & Kann, 2004) are available with which we can carry out this task. But how well do these tools perform?

A study that addressed the question was published by Ahlgren and Kekäläinen (2007). The goal of the research was to find combinations of indexing strategies and query terms that enhances the retrieval of texts written in Swedish. *SWETWOL* was utilized for the strategies that involved compound splitting. The authors concluded that the precision of the software was low when used for this purpose, mainly due to ambiguous compounds; words for which the software generates a list of splitting suggestions.

A study conducted by Sjöbergh and Kann (2004) included the splitting of 3 500 compounds. The authors created a modified version of the spellchecking software *Stava*, and named it *Compound Splitter*. The results show that 99% of all compounds were split and that 97% of them were interpreted correctly. The latter result indicates that the problem of ambiguous compounds was for the most part addressed here, as the software evaluated the options and chose appropriately.

Another related study was carried out by Dalianis (2005). Here, the same compound splitter used by Sjöbergh and Kann (2004) was given the task of splitting 128 Swedish compounds. The main purpose of this research was to investigate the possibility of obtaining search results when using components rather than compounds that originally gave no hits. He presents an accuracy rate of 84% for *Compound Splitter*. The most intriguing part of the results is that the authors managed to obtain 64% more search hits when using the results of the compound splitter.

The benefits of compound splitting can also be traced to clustering according to a research by Rosell (2004). The aim of this study was to improve clustering of Swedish newspaper articles using different techniques, including compound splitting. Again, *Compound Splitter* was the software of choice. The authors recognized a 10 percent improvement in the clustering results by using the compound splitter alone.

1.3 Purpose

While it is clear that compound splitting has been a subject of research for several authors throughout the years, our literature review does highlight a lack of knowledge in this area. Measuring the performance of compound splitting tools in conjunction with words from the field of medicine seems appropriate (Dujols, Aubas, Baylon, & Grémy, 1991). Regarding the choice between the tools at hand, *Compound Splitter* and *SWETWOL*, the clear choice for this study was the first option. Due to time constraints we decided to rule out *SWETWOL*, as the development of an interpretation method would have been necessary for the ambiguous compounds. By measuring the performance of *Compound Splitter* using a set of specialized words, we hope to gain an understanding of its usability in this particular domain.

1.4 Goal and audience

Our goal is to find out how well a general Swedish compound splitter performs on words from the medical world. In order to achieve this goal, the performance of the *Compound Splitter* software will be evaluated. A potential audience of this study is professionals that are attempting to index large amounts of medical texts.

2 Extended background

Some of the related work deserves further discussion. Ahlgren and Kekäläinen (2007) admitted that *SWETWOL* performed well when asked to simply normalize the compound words, for instance, "stålindustrin" (the steel industry) was converted to "stålindustri" (steel industry) correctly. However, when asked to split and normalize compound words, certain examples proved to be ambiguous. The software then produced a list of results where only one was a correct splitting. The compound word "kulturdebatt" (cultural debate) is presented as an example, which generated three suggestions; "kultur|debatt" (culture, debate), "kult|ur|debatt" (cult, arm watch, debate), "kul|tur|debatt" (fun, luck, debate). Using this example among others, the authors argue that the precision of *SWETWOL* could be too low when used to split compounds. This statement is however subject to criticism as it cannot be validated due to the absence of information about the size of the test data set as well as the precision, recall and accuracy of the performed experiments. While the process of compound splitting using *SWETWOL* is a rather small part of the study, these results would have been more credible if more detailed data was presented.

Sjöbergh and Kann (2004) classified 1 300 compounds from their original set of 3 500 as ambiguous. *Compound Splitter* was created by the authors with this problem in mind, as it uses a number of different resources in an effort to identify the correct option. It is important that a certain part of this study is discussed, namely the accuracy figures that are presented. The authors state that "Unless otherwise stated, all accuracy figures in the evaluation are computed only on the ambiguous compounds" (Sjöbergh & Kann, 2004). While the accuracy numbers (99% of compounds were split, 97% of which were correctly interpreted) are impressive, it is worth noting that the test data of concern only included compound words. The absence of non-compound words eliminates the problem of oversplitting words, a term which refers to a split being performed even though it is not necessary, as is the case with words such as "nattliv" (night life). Including such words could very well have decreased the numbers. Another problem that Sjöbergh and Kann also highlight is the lack of an official source with correct compound splits that can be used for comparison with the software output. The authors are forced to use a number of different sources as a substitute, which causes further concerns in regards to the validity of the results. Furthermore, the sample compounds were extracted from a set of 50 000 words in Swedish which originated from the Stockholm-Umeå Corpus (Källgren, Gustafson-Capková, & Hartman, 2006). While the corpus is meant to present a diverse collection of words from many different areas including health and medicine, we do not know to what degree medical words are represented within the data set of 50 000 words.

During his experiments, Dalianis (2005) found that 20 of the 128 compounds were either oversplit, incorrectly split or not split even though they should have been. He concludes that the accuracy rate of *Compound Splitter* in his experiments is 84%. The compound splits conducted by the software are compared to ideal splits specified by the author. The same problem that was discussed in regards to the previous study arises here; one can question the validity of this list of correct splits since no descriptions as to how they were chosen is included in the report.

Despite certain shortcomings, the results from Sjöbergh and Kann (2004) as well as Dalianis (2005) are interesting to compare with the outcome of our own experiments with *Compound Splitter*, since both these studies used the same software but provided different sets of input data.

3 Method

3.1 Choice of method

The aim of our method is to quantify the performance of the *Compound Splitter* software when used on words from the medical domain. Performance in this case is defined as to what degree the software splits compound words correctly and leaves non-compound words un-split.

The method we choose is an evaluation by use of laboratory experiment. When doing laboratory experiments, a setting is created especially for the evaluation of a phenomenon (Boudreau, Gefen, & Straub, 2001). The reason we choose this method is that it allows evaluation of artifacts that are not used in the real world and that a lot of output data can be collected in a short amount of time since the variables are controlled by the experimenters. A drawback of laboratory experiments is that their results are dependent on the created setting and its variables. If the design of the experiment is bad, our results will be of no value.

Another way to evaluate an artifact is by the use of case studies. The artifact is then studied in its real environment for a longer period of time. The advantage of this is that input data is coming from the real world. This was not an option for this study as we do not know of a medical domain where general compound splitting software is used.

As input data to the *Compound Splitter* software, 200 words were extracted from a set of 1 000 000 randomly selected health records in the Stockholm EPR Corpus (Dalianis, Hassel, & Velupillai, 2009). The words chosen are a selection of the most common words that are more than 10 characters long. The reason for the minimum character limit is that otherwise there would be a very low number of compound words. When dealing with words from medical journals there can be some ethical concerns, since such records sometimes contain sensitive information. However, since the words chosen are the most frequently encountered ones, they should not include things like names and phone numbers since those would be fairly unique. We also go through all the words manually to make sure no such information is present. The reason for choosing medical journals as our source was that we believe that they represent a diverse set of medical terms including diseases, symptoms, substances and anatomical names as well as more common words used in the domain. It also made sense to use words from a source that we think could benefit from compound splitting. An alternative source could have been medical books and papers. However, such a source would first need processing to extract an appropriate set of words and we did not have time for this during this study.

To evaluate the software, we also needed correct solutions to compare its results with. The technique used to obtain these was to perform splitting of the above mentioned words manually. The problem with manual classification is that we might split some words wrong, which would skew the results of the software when compared to our incorrect solutions. To alleviate this issue, we will use multiple raters to perform the splits. Another possible way of generating correct answers would be to use some type of software tool. However, this would in effect make our study a software comparison instead of a software evaluation.

To analyze the data, the output from the software is compared with the output from our manual splitting to determine the rate of true and false positives and negatives. A true positive in this case is a compound word that was split correctly. A true negative is a word that is not a compound and was correctly not split by the program. There are two forms of errors we will consider as false positives. The first is compound words that are split in the wrong place, and the second is non-compound words that are split, sometimes referred to as oversplitting. The last classification is false negatives which are compound words that are not split even though they should have been. Some examples of each of these cases are the following:

halsinfektion > hals|infektion	tp (correct split)
medvetande > medvetande	tn (correct not split)
halsinfektion > hal|sinfektion	fp (incorrect split)
medvetande > med|vetande	fp (incorrect split)
halsinfektion > halsinfektion	fn (incorrect not split)

When we have this information, we can calculate accuracy, precision and recall which are commonly used measurements when doing classification tasks. In this context the definitions of these statistics are:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

In this study, accuracy is therefore how correct the software is on a whole since it measures the ratio between the number of correct answers and all answers. The precision rating shows how good the software is at splitting, since it only considers cases were the software tried to do a split and measures how often it was correct. Furthermore, recall is a measure of the software's ability to identify the right compounds and split them correctly. It is sometimes referred to as sensitivity. To calculate the confidence interval of our measures, the following formula was used:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here n is the number of rated words, $\hat{p}$ is the calculated measure (accuracy, precision or recall) and the constant 1.96 is used because the desired confidence interval is 95%.

3.2 Application of method

The first thing we needed to do was to go through our list of words and perform some data cleaning to remove anomalies and make them consistent. The steps of the cleaning process were:

1. Remove any numbers.
2. Remove any blank spaces.
3. Convert all words to lower case.
4. Remove any tags. For example: "<nl>" and "<b>".
5. Remove words containing a slash, since they really are multiple words. For example: "tablett/vecka" (pill per week).
6. Remove words containing a hyphen, since they are already split. For example "lab-status" (laboratory status).
7. Remove words that are names of places. For example "Skärholmen" (a suburban district southwest of Stockholm).
8. Remove words that are shorter than 10 characters.

The next step was to perform the manual splitting. The two of us acted as raters and when we were unsure of the meaning of a word we used search engines. To measure the inter rater agreement we used percentage of agreement. There are also a number of more sophisticated measures like Krippendorff's alpha, Scott's pi and Cohen's kappa which are discussed in for example (Artstein & Poesio, 2008), but we felt a simple approach was sufficient in this case. The first time we performed our rating we noticed a rather low agreement percentage of 76%. Since we were not satisfied with this figure, we constructed two rules to help with the rating of words:

1. The components that the compound is split into must both be proper Swedish words that can be used on their own.
2. The split should not distort the meaning of the compound. This means that a split shall not be performed if one of the components on its own has a meaning that cannot be found in the compound.

When we did the rating again with these rules in mind, we arrived at an agreement of 98% which we found acceptable. This is the figure we will compare the software to, since we can't expect it to perform better than human raters. For the last 2% of the cases where we still disagreed, we decided upon a split after a discussion and defined it as correct. All the manual splits were compiled in a document for later use.

When the manual splitting was completed, the *Compound Splitter* was set up in a Debian Linux environment and we started inputting the same words into the program and noting the results. Finally, we compared these results with our own and classified each split as true positive, true negative, false positive or false negative and calculated a number of measures as explained in the previous chapter.

3.3 Literature study

One of the first tasks in this project was to conduct a literature review to increase our understanding of the area of computational linguistics and specifically compound splitting. The majority of our sources were found online using two search engines for scientific literature; Google Scholar (http://scholar.google.com/) and Web of Science (http://apps.isiknowledge.com/). The table below shows the search queries that were used in conjunction with the different engines and the literature that was located as a result.

Source	Search engine	Search query
Ahlgren & Kekäläinen (2007)	Web of Science	swedish compound splitting
Dujols,Aubas, Babylon & Grémy (1991)	Web of Science	medical compound term
Sjöbergh & Kann (2004)	Google Scholar	swedish compound splitting
Rosell (2004)	Google Scholar	swedish compound splitting
Dalianis (2009)	Google Scholar	swedish compound splitting
Koehn & Knight (2003)	Google Scholar	compound splitting
Macherey, Dai, Talbot, Popat & Och (2011)	Google Scholar	compound splitting
Boudreau, Gefen, & Straub (2001)	Google Scholar	experiment information systems research

Table 1 A number of sources found

In addition, reviewing of the reference lists provided by two of the articles above, that of Ahlgren & Kekäläinen (2007) and Sjöbergh & Kann (2004) resulted in additional sources that were used in this research: Karlsson (Karlsson, 1992) and Källgren, Gustafson-Capková & Hartman (2006).

Using the search queries above, the search engines returned large lists of articles with varying levels of relevance to our research. The process of selecting appropriate sources included several steps and began with an analysis of the article titles in the search results. The articles with titles that included key terms such as "compound splitting" were subject to further analysis where the abstracts were taken into consideration. Using the title and abstract, the articles that clearly were highly relevant to our work were downloaded and stored. A final step was necessary before the articles could be included in our project, namely an extensive review of the described research and the majority of its contents. The methods used; the references and the results were some of the topics that were taken into consideration when evaluating the quality and relevance of a research article.

Finally, the studies by Dalianis, Hassel, & Velupillai (2009) and Artstein & Poesio (2008) were supplied to us by our supervisor.

4 Results

This figure displays the results of the *Compound Splitter* software compared to our manual splitting. Out of the 200 words that were supplied to the software, 161 were correctly handled. This gives an accuracy of 80.5% ± 5.5% with a confidence interval of 95%. This figure can be compared with the inter-rater agreement of the manual splitting which was 98%. In addition, the results show a precision rating of 79.8% ± 5.5% (CI 95%) and a recall rating of 77.17% ± 5.8% (CI 95%).

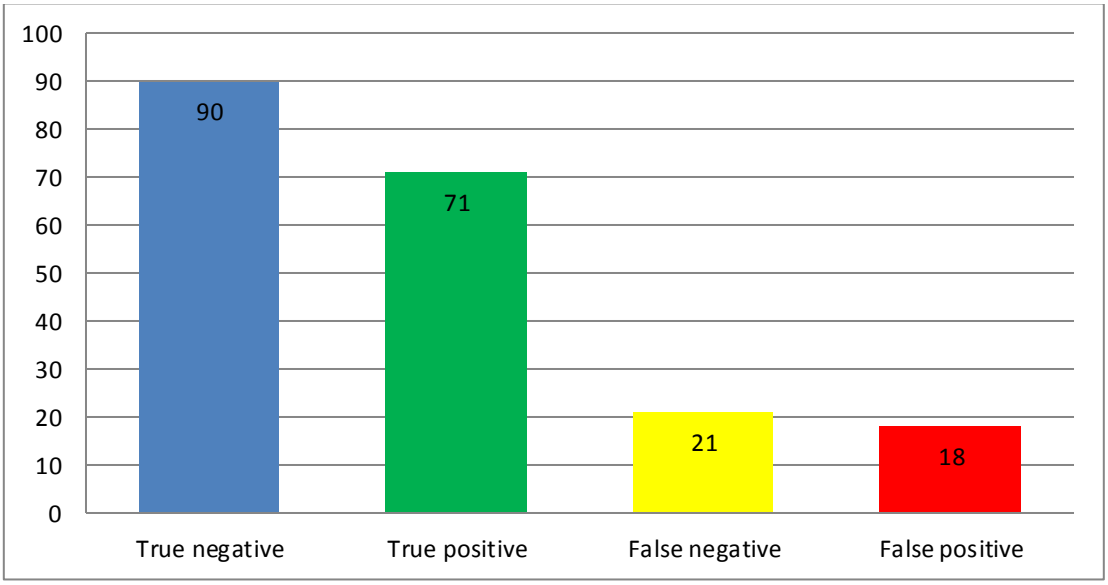

Figure 1 Compound splitter results

The table below contains the final results for a selection of 10 words. The abbreviations in the outcome column have the following meanings: tp=true positive, tn=true negative, fp=false positive and fn=false negative. See appendix A for the full list of words.

Word	Compound Splitter	Manual split	Outcome
kompresser	kompresser	kompresser	tn
bakhuvudet	bak|huvudet	bak|huvudet	tp
femurfraktur	fe|mur|fraktur	femur|fraktur	fp
sannolikhet	sannolikhet	sannolikhet	tn
kompletteras	kompletteras	kompletteras	tn
höftplastik	höft|plastik	höft|plastik	tp
ansträngning	ansträngning	ansträngning	tn
exstirpation	exstirpation	exstirpation	tn
mycostatin	mycostatin	mycostatin	tn
bukaortaaneurysm	bukaortaaneurysm	buk|aorta|aneurysm	fn

Table 2 Sample words from the final results

5 Analysis

As can be seen in the obtained precision of 79.8%, *Compound Splitter* had problems with a number of words. In most of these cases, the software oversplit compound words. An example is the word "reposition" (repositioning of bones) that should not have been split, but was returned as "re|position" by the *Compound Splitter*. This is incorrect as "re" is not a word that can be used independently as is required by our first splitting rule. Another example is the word "femurfraktur" (fracture of the femur bone). The software suggested "fe|mur|fraktur" (fairy, wall, fracture). While all these components can be used on their own, they conflict with our second splitting rule as it distorts the meaning of the original compound.

The recall was slightly lower (77.17%) compared to the precision indicating that the software also failed to identify a number of compounds. Many of these compounds were from the domain of medicine, for example "bukaortaaneurysm" (abdominal aortic aneurysm). The correct split for this compound is "buk|aorta|aneurysm" (abdomen, aorta, aneurysm), but the software returned the original compound without splitting it, probably because it did not recognize the components as individual words. An additional example of this is "smärtstillning" (pain relief) with a correct split being "smärt|stillning" (pain, relief).

Sjöbergh and Kann (2004), the creators of *Compound Splitter* presented a precision of 97% in their study. We suspect that there are several reasons for why the difference in precision is so large. As was expected, the use of medical words proved to be problematic and decreased the overall precision since many of these words are not part of the programs internal dictionary. Another detail that might have had an impact on the results is the difference in test data. While the mentioned study used a set of compound words, our study included a substantial amount of non-compound words. A large amount of the false-positives were such examples that the software incorrectly recognized as compounds.

Dalianis (2005) used *Compound Splitter* in his study and obtained an accuracy of 84%. We can see a small difference in comparison with our number (80.5%). We suspect that this is again related to the test data set. In his studies, Dalianis measured the performance of the software while using 128 words. These words are described as "high frequent Swedish compound queries (6.2 per thousand) with no search results among 1.6 million searches carried out at nine public web sites containing all together 100,000 web pages in Swedish." (Dalianis, 2005). From this description, we can conclude that these words are rather hard to interpret, as the search engines failed to generate any results while using them. This could explain why our results are similar to that of Dalianis. The fact that our numbers are slightly lower could be related to the absence of non-compound words in the test data discussed earlier.

6 Conclusions and discussion

6.1 Conclusions

With the results in hand, we can conclude that *Compound Splitter* with an overall accuracy of 80.5% performs reasonably well, considering that the software is not meant to be used with words from a specific domain such as medicine. However, before it can be used as reliable software for automatic compound splitting in this domain, it should be modified to better handle words of this type.

6.2 Discussion

There are several limitations with our study, mainly related to the methods used. To begin with, the test data set only contained 200 words. This was a deliberate choice due to time constraints, but a larger set of words would have increased the validity of the results.

Another matter that had great impact on the results is the rule set that was used when conducting the manual splitting. These rules, as specified in the methods section were necessary to increase our inter-agreement rate. It must nevertheless be noted that the creation and application of these rules involved some subjective reasoning as there are no standard guidelines for how compound splitting should be performed to produce appropriate components. While the application of the first rule was simple, the second one caused some problems. It is not always clear cut when the meaning of a compound is distorted. Obviously the splitting of "signifikanta" (plural of significant), into "signifik|anta" (significant, assume) is incorrect. Application of this rule is not as simple when considering the split of the word "efterforska" (inquire) into "efter|forska" (after, inquire). Although inquiring after something does not make sense, the Swedish word "efter" can also mean "for" or "about" which together with the word inquiring does not distort the meaning of the original compound. When constructing the rule set, consulting additional experts could have contributed to the overall quality of the rules.

6.3 Future research

An extension of our study could be done on a larger set of words to determine whether our generalizations remain valid. A topic for future research could also be to perform a study on the same set of words but with a different piece of software and compare the results with ours. For example, *SWETWOL* could be used in conjunction with a custom made interpretation method that chooses the correct split. Another thing that could be done is to investigate whether *Compound Splitter* could be modified to work better with a specific domain like that of medicine.

7 Literature references

Ahlgren, P., & Kekäläinen, J. (2007). Indexing strategies for Swedish full text retrieval under different user scenarios. *Information Processing & Management, Vol. 48*(1).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, Vol. 34*(4).

Boudreau, M.-C., Gefen, D., & Straub, D. (2001). Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly, Vol 25*(1).

Dalianis, H. (2005). Improving search engine retrieval using a compound splitter for Swedish. *Proceedings of the 15th NODALIDA Conference.*

Dalianis, H., Hassel, M., & Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and Some Initial Findings. *Proceedings of ISHIMR 2009.*

Dujols, P., Aubas, P., Baylon, C., & Grémy, F. (1991). Morphosemantic Analysis and Translation of Medical Compound Terms. *Methods of Information in Medicine, Vol. 30*(1).

Karlsson, F. (1992). SWETWOL: A Comprehensive Morphological Analyzer for Swedish. *Nordic Journal of Linguistics, Vol. 15*.

Koehn, P., & Knight, K. (2003). Empirical Methods for Compound Splitting. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics.*

Källgren, G., Gustafson-Capková, S., & Hartman, B. (2006). *Manual of the Stockholm Umeå Corpus version 2.0.* Hämtat från Språkbanken: http://spraakbanken.gu.se/eng/resources/suc den 09 12 2011

Macherey, K., Dai, A., Talbot, D., Popat, A., & Och, F. (2011). Language-independent Compound Splitting with Morphological Operations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.*

Rosell, M. (2004). Improving Clustering of Swedish Newspaper Articles using Stemming and Compound Splitting. *Proceedings of the thirteenth ACM international conference on Information and knowledge management.*

Sjöbergh, J., & Kann, V. (2004). Finding the Correct Interpretation of Swedish Compounds - a Statistical Approach. *Proceedings of the 4th International Conference on Language Resources and Evaluation.*

8 Appendix A – Full Results

Word	*Compound Splitter*	Manual split	Outcome
kompresser	kompresser	kompresser	tn
bakhuvudet	bak|huvudet	bak|huvudet	tp
femurfraktur	fe|mur|fraktur	femur|fraktur	fp
sannolikhet	sannolikhet	sannolikhet	tn
kompletteras	kompletteras	kompletteras	tn
höftplastik	höft|plastik	höft|plastik	tp
ansträngning	ansträngning	ansträngning	tn
exstirpation	exstirpation	exstirpation	tn
mycostatin	mycostatin	mycostatin	tn
bukaortaaneurysm	bukaortaaneurysm	buk|aorta|aneurysm	fn
nyupptäckt	nyupptäckt	ny|upptäckt	fn
ytterligare	ytterligare	ytterligare	tn
kortisonbehandling	kortison|behandling	kortison|behandling	tp
tillkommit	tillkommit	till|kommit	fn
resttillstånd	rest|tillstånd	rest|tillstånd	tp
reponering	re|ponering	reponering	fp
insulatard	insulatard	insulatard	tn
vårdplanering	vård|planering	vård|planering	tp
energibehov	energi|behov	energi|behov	tp
någonstans	någonstans	någonstans	tn
besvärliga	besvärliga	besvärliga	tn
tillstöter	till|stöter	till|stöter	tp
underarmen	under|armen	under|armen	tp
buksmärtorna	buk|smärtorna	buk|smärtorna	tp
kontrollerats	kontrollerats	kontrollerats	tn
vätskeintag	vätske|intag	vätske|intag	tp
nydiagnostiserad	ny|diagnostiserad	ny|diagnostiserad	tp
normaliserats	normaliserats	normaliserats	tn
trombocytopeni	trombocytopeni	trombo|cytopeni	fn
psykosociala	psyko|sociala	psyko|sociala	tp
radiologiskt	radiologiskt	radiologiskt	tn
polymyalgia	polymyalgia	polymyalgia	tn
ryggsmärta	rygg|smärta	rygg|smärta	tp
omständigheter	omständigheter	omständigheter	tn
habiliteringen	habiliteringen	habiliteringen	tn
ostadighet	ostadighet	ostadighet	tn
skriftliga	skriftliga	skriftliga	tn
ångestproblematik	ångest|problematik	ångest|problematik	tp
irritation	irritation	irritation	tn
Riskfaktorer	risk|faktorer	risk|faktorer	tp

Word	*Compound Splitter*	Manual split	Outcome
observation	observation	observation	tn
analgetika	analgetika	analgetika	tn
antireumatika	antireumatika	antireumatika	tn
sårkanterna	sår|kanterna	sår|kanterna	tp
symmetrisk	symmetrisk	symmetrisk	tn
obehandlad	obehandlad	obehandlad	tn
antibiotika	antibiotika	antibiotika	tn
oförändrad	oförändrad	oförändrad	tn
fortsättning	fortsättning	fortsättning	tn
instruktion	instruktion	instruktion	tn
resterande	resterande	resterande	tn
sårinfektion	sår|infektion	sår|infektion	tp
erfarenhet	erfarenhet	erfarenhet	tn
begynnande	begynnande	begynnande	tn
labbmässigt	labbmässigt	labbmässigt	tn
telefonledes	telefon|ledes	telefonledes	fp
överviktig	över|viktig	över|viktig	tp
efterförlopp	efter|förlopp	efter|förlopp	tp
komplement	komplement	komplement	tn
poliklinisk	poli|klinisk	poliklinisk	fp
insulinbehandling	insulin|behandling	insulin|behandling	tp
morgonstelhet	morgon|stelhet	morgon|stelhet	tp
överenskommer	överens|kommer	överens|komma	tp
appendicit	appendicit	appendicit	tn
accepteras	accepteras	accepteras	tn
kombination	kombination	kombination	tn
rektalcancer	rek|tal|cancer	rektal|cancer	fp
illaluktande	illa|luktande	illa|luktande	tp
förändringar	förändringar	förändringar	tn
ansträngningsutlöst	ansträngnings|utlöst	ansträngnings|utlöst	tp
välfungerande	väl|fungerande	väl|fungerande	tp
hemsjukvård	hem|sjukvård	hem|sjuk|vård	fp
komplettera	komplettera	komplettera	tn
stödkontakt	stöd|kontakt	stöd|kontakt	tp
gångsvårigheter	gång|svårigheter	gång|svårigheter	tp
insjuknande	insjuknande	insjuknande	tn
undersökning	under|sökning	under|sökning	tp
smärtstillning	smärtstillning	smärt|stillning	fn
anslutning	anslutning	anslutning	tn
avlastning	avlastning	av|lastning	fn

Word	*Compound Splitter*	Manual split	Outcome
sedvanliga	sedvanliga	sed|vanliga	fn
utvecklingen	utvecklingen	utvecklingen	tn
pacemakersystem	pacemaker|system	pacemaker|system	tp
slutgiltiga	slut|giltiga	slut|giltiga	tp
arbetsträning	arbets|träning	arbets|träning	tp
exempelvis	exempelvis	exempelvis	tn
telefonsamtal	telefon|samtal	telefon|samtal	tp
viktstabil	vikt|stabil	vikt|stabil	tp
väntelista	vänte|lista	vänte|lista	tp
telefonuppföljning	telefon|uppföljning	telefon|uppföljning	tp
hjärtfrekvensen	hjärt|frekvensen	hjärt|frekvensen	tp
ögondroppar	ögon|droppar	ögon|droppar	tp
eftermiddag	eftermiddag	eftermiddag	tn
antikoagulantia	antikoagulantia	antikoagulantia	tn
kärlkirurgen	kärl|kirurgen	kärl|kirurgen	tp
symtomgivande	symtom|givande	symtom|givande	tp
komplicerat	komplicerat	komplicerat	tn
obstruktivitet	obstruktivitet	obstruktivitet	tn
imorgonbitti	imorgonbitti	imorgon|bitti	fn
konsulterar	konsulterar	konsulterar	tn
amputation	amputation	amputation	tn
radiusfraktur	radiusfraktur	radius|fraktur	fn
tunntarmen	tunn|tarmen	tunn|tarmen	tp
begynnande	begynnande	begynnande	tn
förbättras	förbättras	förbättras	tn
läkemedelslistan	läkemedels|listan	läkemedels|listan	tp
kreastegring	kreastegring	kreastegring	tn
hjälpsökande	hjälp|sökande	hjälp|sökande	tp
anmärkning	anmärkning	anmärkning	tn
röntgenologisk	röntgenologisk	röntgenologisk	tn
uppmärksam	uppmärksam	uppmärksam	tn
underlätta	under|lätta	underlätta	fp
medicinerat	medicinerat	medicinerat	tn
sårmottagning	sår|mottagning	sår|mottagning	tp
reumatalog	reumatalog	reumatalog	tn
falltendens	fall|tendens	fall|tendens	tp
förutsättningar	förut|sättningar	förutsättningar	fp
andningspåverkad	andnings|påverkad	andnings|påverkad	tp
muskelvärk	muskel|värk	muskel|värk	tp
accepterar	accepterar	accepterar	tn

Word	*Compound Splitter*	Manual split	Outcome
kontrollerat	kontrollerat	kontrollerat	tn
färdigbehandlad	färdig|behandlad	färdig|behandlad	tp
remitterad	remitterad	remitterad	tn
primärvården	primär|vården	primär|vården	tp
signifikanta	signifik|anta	signifikanta	fp
habitualtillstånd	habitualtillstånd	habitual|tillstånd	fn
protokollet	protokollet	protokollet	tn
möjligheten	möjligheten	möjligheten	tn
otosalpingit	otosalpingit	otosalpingit	tn
knäplastik	knä|plastik	knä|plastik	tp
biverkningar	biverkningar	biverkningar	tn
natriumbikarbonat	natrium|bikarbonat	natrium|bikarbonat	tp
hörselkontroll	hörsel|kontroll	hörsel|kontroll	tp
avvaktande	avvaktande	avvaktande	tn
informerar	informerar	informerar	tn
oförändrade	oförändrade	oförändrade	tn
behandlingsindikation	behandlings|indikation	behandlings|indikation	tp
avtackling	avtackling	avtackling	tn
medicinjustering	medicin|justering	medicin|justering	tp
cholecystit	cholecystit	cholecystit	tn
uppresning	uppresning	upp|resning	fn
svårigheter	svårigheter	svårigheter	tn
vårdplaneras	vård|planeras	vård|planeras	tp
mastektomi	mastektomi	mastektomi	tn
misstänkta	miss|tänkta	misstänkta	fp
muskelsvaghet	muskel|svaghet	muskel|svaghet	tp
receptförnyelse	recept|förnyelse	recept|förnyelse	tp
koronarangiografi	koronarangiografi	koronar|angiografi	fn
divertikulit	divertikulit	divertikulit	tn
omfattning	omfattning	omfattning	tn
cirkulation	cirkulation	cirkulation	tn
sjukgymnastik	sjuk|gymnastik	sjuk|gymnastik	tp
tillfälligt	tillfälligt	tillfälligt	tn
viktnedgången	vikt|ned|gången	vikt|nedgången	fp
utgjutning	utgjutning	ut|gjutning	fn
vänstersidig	vänstersidig	vänster|sidig	fn
tonsillectomi	tonsillectomi	tonsill|ectomi	fn
sårkontroll	sår|kontroll	sår|kontroll	tp
omväxlande	omväxlande	omväxlande	tn
adenocarcinom	adenocarcinom	adenocarcinom	tn

Word	*Compound Splitter*	Manual split	Outcome
cystoskopi	cystoskopi	cystoskopi	tn
övervakning	övervakning	över|vakning	fn
hörapparat	hör|apparat	hör|apparat	tp
iohexolclearance	iohexolclearance	iohexol|clearance	fn
detaljerad	detaljerad	detaljerad	tn
smärtlindrande	smärt|lindrande	smärt|lindrande	tp
lipidstatus	lipidstatus	lipid|status	fn
reservation	reservation	reservation	tn
fortsätter	fort|sätt|ter	fortsätter	fp
gränsvärde	gräns|värde	gräns|värde	tp
överenskommer	överens|kommer	överens|kommer	tp
intresserad	intresserad	intresserad	tn
gynekologisk	gynekologisk	gynekologisk	tn
återkommer	återkommer	åter|kommer	fn
synnedsättning	syn|ned|sättning	syn|nedsättning	fp
återbesöket	åter|besöket	åter|besöket	tp
påputtning	på|puttning	påputtning	fp
medföljande	med|följande	med|följande	tp
reposition	re|position	reposition	fp
tablettbehandling	tablett|behandling	tablett|behandling	tp
borttagande	bort|tagande	bort|tagande	tp
uroterapeut	uroterapeut	uroterapeut	tn
förskriver	för|skriver	förskriver	fp
efterforska	efter|forska	efter|forska	tp
blodtryckskontroll	blod|tryck|s|kontroll	blod|trycks|kontroll	fp
långvariga	lång|variga	lång|variga	tp
instruktioner	instruktioner	instruktioner	tn
obstruktiva	obstruktiva	obstruktiva	tn
recidiverande	recidiverande	recidiverande	tn
faecesodling	faecesodling	faeces|odling	fn
accentuerade	accentuerade	accentuerade	tn
näsfraktur	näs|fraktur	näs|fraktur	tp
kommunikation	kommunikation	kommunikation	tn
medicinjusteringar	medicin|justeringar	medicin|justeringar	tp
respektive	respektive	respektive	tn
vänsterkammarfunktion	vänster|kammar|funktion	vänster|kammar|funktion	tp
acetylcystein	acetylcystein	acetyl|cystein	fn
flebografi	flebografi	flebografi	tn
biventrikulär	bi|ven|tri|kul|är	biventrikulär	fp
behandlingen	behandlingen	behandlingen	tn

Stockholm University
Department of Computer and Systems Sciences
16440 Kista
Telephone: 08–16 20 00
http://dsv.su.se