Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus

PONTUS STENETORP

KTH Computer Science
and Communication

Master of Science Thesis
Stockholm, Sweden 2010

Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus

PONTUS STENETORP

Master's Thesis in Computer Science (30 ECTS credits)
at the School of Computer Science and Engineering
Royal Institute of Technology year 2010
Supervisor at CSC was Viggo Kann
Examiner was Stefan Arnborg

TRITA-CSC-E 2010:017
ISRN-KTH/CSC/E--10/017--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication

KTH CSC
SE-100 44 Stockholm, Sweden

URL: www.kth.se/csc

Abstract

This thesis presents an automated system for extracting neologisms using machine learning approaches. The neologisms are extracted from a large temporally annotated corpus containing newspaper articles and blog posts. We find that our system is different from much of the previous research on neologism extraction and justify these differences by relating it to current research in evolutionary linguistics.

Our main contribution is a system which can incorporate a larger amount of features than any previous system. Our approach also enables multiple annotators to provide feedback to the system and affect the way the system ranks words in regard to “newness”.

In addition, our system is capable of assisting users manually extracting neologisms by providing statistics regarding word usage over time. We also present and analyze words annotated by twelve anonymous annotators in regard to their “newness” and are able to draw conclusions on how language users perceive neologisms.

Referat

Automatisk excerpering av svenska nyord med en tidsannoterad korpus

I denna rapport presenterar vi ett system som med hjälp av metoder inom maskininlärning automatiskt kan extrahera nyord. Nyorden extraheras från en stor tidsannoterad corpus som innehåller texter från tidningar och bloggar. Vi finner att systemet skiljer sig från tidigare konstruerade system för samma syfte och rättfärdigar detta genom att sätta systemet i relation till forskning inom evolutionärlingvistik.

Resultatet blir ett system som kan ta hänsyn till en större mängd ordattribut än något tidigare system. Vår ansats möjliggör även att fler än en individ kan bidra till hur systemet klassificerar ord med avseende på ”nyhetsgrad”.

Utöver detta kan systemet hjälpa dess användare att manuellt extrahera nyord genom att presentera statistik gällande ordens användning över en given tidsperiod. Vi presenterar och analyserar även ord annoterade av tolv anonyma annoterare med avseende på ”nyhetsgrad” och kan dra slutsatser från detta om hur språkanvändare ser på fenomenet nyord.

Contents

1 Introduction 1

1.1 Purpose . . . 2

1.2 Assumptions, Definitions and Limitations . . . 2

1.3 Disposition . . . 3

I Linguistic Theory and Previous Research 5

2 Linguistic Implications 7

2.1 Defining Language Borders . . . 7

2.2 Defining Neologisms . . . 9

2.3 Summary . . . 11

3 Related Research 13

3.1 Methods of Neologism Extraction . . . 13

3.1.1 Determining Lexical Borrowing . . . 13

3.1.2 Extraction Using Lexical Cues . . . 14

3.2 Systems for Neologism Extraction . . . 14

3.3 Summary . . . 15

II Constructing a Novel System for Extracting Neologisms 17

4 Constructing a Temporally Annotated Corpus 19

5 An Active System for Extracting Neologisms 23

5.1 Feedback and Result Cycle . . . 23

5.2 Word Features . . . 25

5.3 User Interface . . . 26

IIIExperiments and Results 29

6 Training Data 31

6.1 Acquiring Initial Training Data 31
6.2 Acquiring Diverse Training Data 33
6.3 Analysis of Annotated Data 35
6.4 Summary 36

7 Results **39**
7.1 Aiding Lexicographic Extraction of Neologisms 39
7.2 Automatic Extraction of Neologisms 39
7.3 Results Compared to the Annual List of Neologisms of 2009 40

8 Conclusions **45**
8.1 Fulfillment of Purpose 45
8.2 Future Work 45

Bibliography **49**

Appendices **51**

A Svenska Nyord Corpus XML-format **53**

B Training Data Annotation Application User-interface **55**

C Additional Annotation Data Analysis **57**

Glossary **61**

Chapter 1

Introduction

Languages are ever-changing entities, new words arise and others are forgotten. Some words are even borrowed between languages and blurs the borders between what is usually refered to as different languages. It is well known that even grammar can be borrowed between languages, but this thesis focuses on the creation and borrowing of neologisms or informally, *"new words"*.

Both language-users and language researchers have an interest in tracking and noting changes to a language. A lexicographer is likely to be interested in both how established a certain word is, how it is commonly used and the origin of the word. Is the word borrowed from another language? Where did it first arise? Linguists on the other hand are interested in observing patterns and acquiring examples supporting their frameworks and theories.

An ordinary language-user is likely to be more practical in his or her need for knowledge in regard to neologisms. Most likely he or she is mostly concerned with understanding a specific text and does not wish to be stumped when encountering a word that he or she does not recognize. A language learner encountering a word that is yet to be noted in mainstream dictionaries is likely to be lost and left to rely on context or other language-users already proficient in the language.

For the Swedish language there are two main sources for neologisms. First there is the de-facto dictionary of modern Swedish, The Dictionary of the Swedish Academy[1] (Svenska Akademien, 2006) "Svenska Akademiens Ordlista", which for each new edition provides a section on new words incorporated into the dictionary since the last edition. There is also the annual list of neologisms published by the Swedish Language Council "Svenska Språkrådet" in the language magazine "Språktidningen" and on the Internet[2]. The Swedish Language Council has also released two books on neologisms, Svenska Språkrådet (2000) and Svenska Språkrådet (2001).

To date, all extraction of neologisms is largely carried out by human efforts. The Swedish Language Council employs eight volonteers who are trained and each

[1] `http://www.svenskaakademien.se/web/Svenska_Akademiens_ordbok.aspx` *(in Swedish)*

[2] `http://www.sprakradet.se/nyord` *(in Swedish)*

assigned a daily newspaper (Lindgren, 2007). As they notice new words in the newspaper they note them and submit their results on a monthly basis to be collected by the language council and serves as a basis for the annual list of neologisms.

1.1 Purpose

The purpose of this thesis is to investigate the possibility of automating the process of neologisms extraction using Natural Language Processing (NLP) techniques. Then to use the approach which is most suitable to construct a system capable of assisting a user in finding current neologisms in the Swedish language.

It is not believed that any automated system can fully replace human efforts, but the thesis should aim to present a system which eases the time-consuming task of reading large amounts of text and manually noting new words. This task can be seen as two-fold, providing suggestions to a human about which words that may be potential neologisms and also to provide metrics which can be used to judge any given word. Currently the Swedish Language Council mainly employs occurences and the abstract notion of novelty, to judge a word in regard to inclusion in the annual list of neologisms. While these methods have proven sufficient, employing automated methods has the potential of improving the amount of words considered for inclusion and minimizing the amount of human effort.

1.2 Assumptions, Definitions and Limitations

In this thesis we assume a very simple notion of what is considered to be a lexeme or informally a *"word"*, we will also make use of the word "lexeme" and "word" interchangeably.

Definition 1 (Lexeme) *A lexeme is:*
Any space or slash delimited string that is stripped of any preceeding and trailing punctuation character unless the punctuation character is connected with another punctuation character of the same type which is not preceeding or trailing.

Since the definition at first may seem slightly complex the reader in encouraged to take a glance at Table 1.1 which illustrates the implications of this definition. Readers familiar with computer-based tokenization systems will see that our definition works closely with what such a system produces when applied to a text.

Capturing the notion of semantics is considered more difficult than considering only lexemes. In accordance to previous work on neologism extraction, which we will discuss further in Part I, we do not take semantical changes to lexemes into consideration (e.g. recently the word "tweet" has also acquired the meaning of "to post a message on a micro-blog", this is an example of a semantical change in meaning). Only lexical borrowing and the creation of new lexemes are considered. We do however propose extensions to our approach in order to capture semantic changes to a word, this is done in Section 8.2 in relation to future work.

String	Is a Lexeme?
Him	Yes
Him...	No
Him/Her	No
Ex-husband	Yes
“Him”	No
“Ex”-husband	Yes

Table 1.1. Several strings classified according to Definition 1

While neologisms is a phenomena that can be tracked continously over time, all the data and examples in this thesis are limited to using data up until and including the 31st of December 2009. The reason for this limitation is to make sure that tests and comparisons are carried out on the same data and also to be able to compare the final results to the Swedish Language Council’s annual list of neologisms from 2009. As will be discussed in Section 4, the corpus used in this thesis contains additional data after the date chosen as a limit for this thesis.

The language discussed and on which any methods are evaluated is the Swedish language. While this is a limitation it is the opinion of the author that the methods proposed in this thesis can be applied to any written natural language. Extensions of the current approaches for other languages are discussed in Section 8.2.

1.3 Disposition

In Part I we first discuss the underlying linguistic theory related to neologisms. In the same part we also discuss previous work on the extraction of neologisms and their results. We then proceed to Part II where we turn to the design and methods for constructing our system. First we discuss the collection and composition of our corpus, then we explain the details of our system for neologism extraction.

For Part III we acquire tranining data for our system by exploiting previously publisted lists of neologisms and by manual annotation of words carried out by human annotators. We also discuss the results of the human annotation and how it relates to the performance of our system and linguistic theory. In conclusion we discuss the performance of our system and how it can aid users when extracting neologisms.

Part I

Linguistic Theory and Previous Research

Chapter 2

Linguistic Implications

This chapter discusses the linguistic foundations for neologisms and language change. First we discuss the definition of languages. Then we proceed to discussing different views on neologisms and what implications our linguistic view on languages has when deciding on a definition of neologisms which is suitable to serve as a basis for a neologism extraction system.

2.1 Defining Language Borders

When discussing neologisms it is necessary to consider a language as a whole. We all have some sense of what a language is, you are aware of when you are speaking in English and when speaking in German. You can distinguish Italian writing from Arabic writing simply by observing the characters. But these languages are highly associated with a country or a culture and once we observe more local differences we can see that what at first may feel intuitive and obvious becomes vague and arbitary.

This criticism of languages as a concept is brought up by Croft (2000) as an issue when observing language change. Croft argues that linguists tend to idealize languages and create an abstract system which can be labeled as being a language. Reality on the other hand is not ideal and poses issues for such an abstract system. Languages share many attributes, these attributes can be due to cultural exposure or the sharing of a common ancestral language. How can we then separate different languages? Also, how can we then determine when a lexeme is new if there is no absolute factual point in time when the word was introduced to the language?

This problem of defining a language has many things in common with the biological problem of defining species. You can easily see that a cat and a dog are different, but they too share common attributes and on a molecular level a large portion of their genetic code. This problem of telling species apart goes back several hundred years. At first breeding was proposed as a definition, species are animals which can breed with each-other. But even this definition falls short from capturing reality since as has been mentioned by Dennett (1995) there are species which defies even this definition and form breeding patterns such as those illustrated in Figure 2.1.

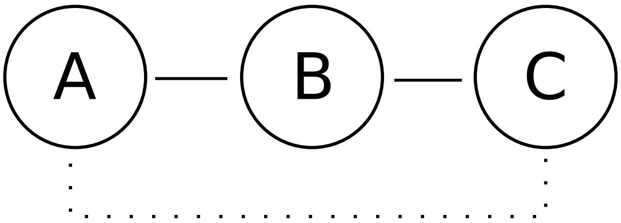


Figure 2.1. A simple illustration of the example given in (Dennett, 1995). Three inviduals from different geographic locations *A*, *B* and *C*. *A* can breed with *B*, *B* can breed with *C*, but *C* can not breed with *A*

The same anology can be applied to communication using a common language. Historically the language which we consider to be modern Swedish must have been gradually changed over time from some ancestral language into modern Swedish. If we look back on texts produced a few generations ago, we can read them without a great deal of effort. But as we go further back in time it becomes gradually more difficult to understand the text as key components of our modern language disappear or are transformed into forms unknown to us. Given this fact, when can we say that Swedish came into existence? And when did it stop being the language it was before it was considered to be Swedish? Assuming that we had a perfect record of a language over time, we would still argue that answering these two questions may very well be impossible.

Accepting this notion of a language we can never in a formal manner fully define what is and what is not a specific language. Instead we are left with a notion which is not easily transfered to a computer and can thus not be used when constructing methods for neologism extraction. But we can safely assume that every language speaker has some notion of what they personally considered to be a certain language. It is however doubtful that all of these notions coincide, but that there is a great degree of overlap between these views, a concept which is illustrated in Figure 2.2.

What we finally arrive at is a notion of a language which is largely statistical, but is based on something from the outside world rather than an idealized abstract notion. Both definitions are difficult to incorporate into a computer-based system, but looking at how language is perceived and utilized by users grants us the possiblity of a more applicable approach since we can rely on data we collect from language users. Instead of attempting to construct an abstract language model by employing experts in linguistics to decide on what is and what is not a part of a language.

While this may not appear to be humble towards much of what has largely been the product of linguistic research, we should be reminded that we may assume different perspectives on an issue. We can once again draw a link to biology, as proposed by Dawkins (1982) it is possible to view organisms on several levels. One

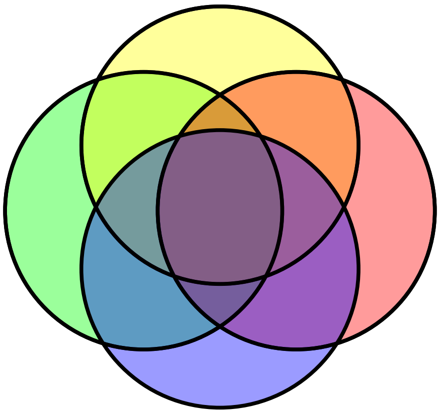


Figure 2.2. Four language users have their views on what is considered to be and not to be their language. On some matters they agree and the circles intersect, on other matters they disagree and there is no intersection

may view rabbits as a species or view the rabbits as an effect of their genetic structure. We have both a species-centric view and a gene-centric view, switching between the two can help us adapt to new situations and give us further insight.

When compiling a book on grammar you should hardly try to view language as a whole and base your grammar on collected statistics on language usage, for this kind of research the notion of an idealized abstract language is suitable. But when observing language change we should not limit ourselves to changes to an abstract language, we should then assume another view which is not more correct of a view than the other, but enables us to view languages on a different level and base our conclusions on direct observations of how a language is applied in real-world situations.

2.2 Defining Neologisms

Defining neologisms may at first appear to be trivial. But as is pointed out by Rey (1995), saying that a word is a neologism is an expression of a subjective opinion rather than something factual and tangible. Once again we are faced with the same issue as we faced in regard to languages in Section 2.1 and are unable to acquire a formal definition of what a neologism is.

Four definitions are proposed by Cabré (1999), while not being definite they largely capture how neologisms have been perceived when constructing systems for extracting neologisms. According to Janssen (2005b) previous systems for neologism extraction tends to assume an approach similar to one of Cabrés definitions without any intricate discussion or justification. We will now introduce these definitions and discuss them briefly.

Definition 2 (Neologism by Diacrony) *A neologism is:*

a [lexeme] that has arisen recently.

At first glance the definition of diacrony is adequate when capturing what is and what is not a neologism. But as we consider several corner-cases we discover that it is perhaps too inclusive, what about misspellings and words which are introduced in a small part of the language community as jargon? Existence does not necessarily imply relevance and novelty.

Definition 3 (Neologism by Lexicography) *A neologism is:*
a [lexeme] that is not present in dictionaries.

Most languages have one or several de-facto standard dictionaries, thus it would be tempting to consider such a dictionary as a source of which lexemes that currently are considered to belong to the language. But dictionaries has several short-comings, they are by their very definition incomplete. They can not possibly include all word-forms. They also have a considerable inclusion bias since they are compiled by few language users. Dictionaries always lag behind the lexical zeitgeist since a dictionary will always contain out-of-date words which will stump many language users and fail to include the very latest of words.

The most serious issue however, is the same as is mentioned for the definition based on diacrony. How do we separate new words from jargon? How can we know that a word is well-established in the language community? A dictionary gives no guidelines on such a matter.

Definition 4 (Neologism by Systematic Instability) *A neologism is:*
a [lexeme] that exhibit signs of formal instability (e.g. morphological, graphic, phonetic) or semantic instability.

Defining neologisms by systematic instability nicely captures many aspects of what is a neologism. Although we have stated that we disregard of the semantics of a word, this definition does capture semantical changes to a word. A problem when accounting for changes in morphology is that it assumes that we have some way of tracking a word. An example from Swedish is “okej” from the English “okay”, it is commonly shortened to “ok” which carries the same semantical meaning but in a different form. Even more troubling is that there is already a Swedish word “ok” meaning “yoke”. Constructing a system capable of separating these meanings and tracking how they change over time is a task worthy enough of a thesis on its own. It therefore seems implausible to base any system on this notion.

Definition 5 (Neologism by Psychology) *A neologism is:*
a [lexeme] that speakers perceive as being a new [lexeme].

The last definition reasons in a way which is similar to how we reasoned regarding languages in the previous section. We can expect a language user to have a sense of what is new and what is not. But as pointed out by Rey (1995), even in this case the language user is sometimes “wrong” and may be fooled by the structure of a word into believing that it may be foreign and borrowed while in fact it is the artefact of a borrowing taking place hundreds of years ago. But is that really wrong? If a word is forgotten and brought back into active use, is it then not new to the language users of that time? Although promising we can never hope that a computer-based system can be on-par with a language user when passing judgements on the inherent “newness” of words.

None of the mentioned definitions are exclusive of one and another, instead they reflect different aspects of what may indicate that a word is indeed a neologism. It is then the case that any method for neologism extraction should seek to incorporate as many as possible of the aspects which are relevant for defining a new word. Keeping in mind Definition 5 and the initial comment in Rey (1995), what is a neologism is largely a subjective notion that is left to a language user to decide.

We can assume that just as language speakers have different opinions on what is and what is not their language as shown in Figure 2.2, different language users will differ in what they consider to be a neologism. Their opinion must be based in some sort of reality that at least partially is captured by the definitions presented in Cabré (1999), possibly even more aspects than those included in the definitions must be considered. In the coming chapter we will discuss how the previous research on neologism extraction mirrors the underlying theory.

2.3 Summary

We have discussed the definition of a language since it is essential when defining what is a new word in a language. Doing so we arrived at the conclusion that when observing language change it is difficult to find a formal definition of what is and what is not a certain language. Instead we are left with a notion based on statistics and observing actual language usage in real-world situations.

We then discussed the definitions used when defining neologisms and how they relate to our previous discussion regarding languages. We found that a neologism carries many aspects which have to be considered if we are to judge a word as worthy of being considered as new or not. We were unable to arrive at a formal definition, but assumed that a word in its nature somehow carries different aspects which human language users can rely on when judging whether they consider a word to be new or not.

Chapter 3

Related Research

The task of neologism extraction has received little attention by researchers. Although there have been several systems constructed for various languages, there is little or no unification in the efforts undertaken by the various groups. Possibly this is due to the task being mainly an interest for lexicographers, which tend to focus on a single language and be employed by publishing companies. In academia it is likely that the lack of metrics and the problems of de-marcation mentioned in the previous chapter are discouraging since it is difficult, if not impossible, to compare the quality of two systems.

In this chapter we will introduce and summarize the efforts that have been undertaken to approach the problem of classifying and extracting neologisms. We begin by introducing several methods for neologism extraction and then proceed to present several full-scale systems that aim to assist and automate the task.

3.1 Methods of Neologism Extraction

Methods for neologism extraction may be employed as a basis for a system extracting neologisms. The research here presents and evaluates approaches which may be essential when constructing such a system.

3.1.1 Determining Lexical Borrowing

As discussed in Section 2.2, a system which can track words or determine their origin would prove to be beneficial when determining the "newness" of a word. Due to language political reasons research in this area has mainly been concerned with anglicisms.

How easy one may detect anglicisms in a language is of course closely related to the structure of the language itself. Detecting English influence on a word in Japanese for example is rather trivial since it is commonly spelled out in the "katkana" alphabet. For example "ball" becomes "ボール" or "booru" when transcribed into Roman letters, if the word was native it would be most likely be represented using “kanji” ideograms or the “hiragana” alphabet instead. When looking into previous research it is thus important that the language which is the subject of the research is closely related to Swedish. Doing this we are forced to disregard of our notion of language independence for a short while.

Precisely due to language political reasons the Norwegian Language Council has carried out research on determining whether or not a word is an anglicism (Andersen, 2005). Research carried out on Norwegian should be considered applicable to Swedish since they are closely related. What was found was that it is difficult to determine whether or not a word is of English origin using relatively simple methods. Reasons for this may be due to the relatedness of Norwegian and English, which would also have implications for Swedish. Similar results were found by Cartoni (2006) in regard to Italian anglicisms.

3.1.2 Extraction Using Lexical Cues

Intuitively, when introducing a new word to a reader you are likely to try to leave clues to the reader that the word is somehow different. Under the assumption that neologisms are more likely to be preceeded by certain expressions (e.g. “so called” or “known as”) along with lexical cues such as quotation-marks, (Paryzek, 2008) succeeded in extracting a large list of neologisms from the scientific magazine “Nature”. The method was largely successful, but it can be argued that scientific texts, are differently structured compared to other texts. Despie this the method shows great potential and is no longer just an intuitive notion but a proven attribute for neologisms.

A method using lexical cues is, unlike the research on anglicisms in Section 3.1.1, largely language independent. This is yet another aspect speaking in favour of employing lexical cues for the extraction of neologisms.

3.2 Systems for Neologism Extraction

For Portugese, Janssen (2005a) used a combination of the definitions proposed by Cabré (1999) to construct a system using both a lexicon and corpus to classify neologisms. Janssen argued that filtering must be kept to a minimum in order to avoid losses of words which are in fact neologisms. The system took a text as input and checked each word towards a lexicon and an old corpus. It then presented the words as potential neologisms to the user if it was not present in either the corpus or the lexicon. If a word was marked by the annotator as a neologism it was incorporated into the internal lexicon and thus not presented as a neologism for future texts.

For the nordic languages there has been several systems constructed to extract neologisms. These systems deserve a high degree of attention due to their relation to the Swedish language.

For Danish a system using search-logs from an online dictionary in combination with a newspaper corpus and an online youth-magazine has been constructed and presented by Halskov (2007). The author raises the question on how to capture sematical changes to the meanings of words and also that more efforts needs to be put into filtering out non-neologisms from the results.

For Icelandic there exists a system described by Bjarnadóttir and Rögnvaldsson (2007) which takes any Icelandic text as input and runs it towards a morphological dictionary and informs the users which words that previously did not exist in the dictionary. Filtering of foreign words was proposed as an improvement, along with lemmatizaton was suggested as further improvements to the system.

For Norwegian the existing system is word list based (Hofland, 2007), no filters are applied but the system employs a novel classification algorithm which allows the words to be sorted into categories of neologisms. Unlike the systems previously mentioned, the Norwegian system is an active system that is updated on a daily basis by incorporating texts from an online newspaper corpora.

3.3 Summary

We find relatively few examples of previous systems and the ones we find employ a wide range of approaches. Filtering out non-neologisms, if done at all, is done by using simple rules according to word structure or statistics such as frequencies. A major issue is the high degree of noise introduced by misspellings and well-established words since both approaches based on lexicons and corpora have a limited coverage.

To remedy this the researchers propose collection of more data, better and more elaborate filtering and statistics on a word-per-word basis. But as we have argued in regard to languages and neologisms, evaluation and classification may be hard and for very good reasons. Still, word attributes such as lexical cues show positive results, although for a very limited domain.

Part II

Constructing a Novel System for Extracting Neologisms

Chapter 4

Constructing a Temporally Annotated Corpus

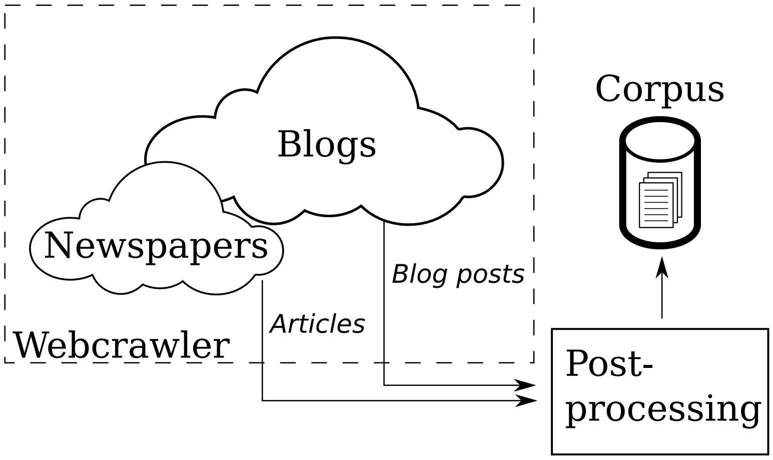


Figure 4.1. System automatical corpus construction from the web

Most Swedish corpora have primarily been collected to analyze Swedish as a language. For this purpose it is not necessary to acquire texts that are temporally annotated. Rather the text annotations will be in the form of part-of-speech and other grammatical annotations. It is not uncommon for these corpora to be more than a century old and that makes them a poor source of neologisms.

With the advent of the Internet, many newspapers started publishing their articles online. This attracted the attention of NLP researchers who saw potential in using Internet resources as a corpus, Hofland (1998) and Hassel (2001). The data collected could then be used to analyze the flow of newspaper texts and track current events, as done in (Eiken et al., 2006).

However using the Internet as a corpus is not problem free. The online texts are not delivered in a clean format, but require some process to extract the actual text from the web document or we risk introducing noise or perhaps even autonyms (McFedries, 2003). There is also the issue of copyright. In accordance with Swedish copyright law, any statistics or excerpts of data protected by copyright is not covered by the copyright of the original text. It thus seems feasible that we could extract a large corpus by employing relatively simple methods and acquire texts from the Internet.

We construct our system for corpus construction to collect newspaper texts on a daily basis from two major Swedish newspapers and from a selection of blogs. The first run of the collector extracts the newspaper archive of each newspaper in order to enable us to have a wide time-span. This is essential since we are aiming at collection data on how the Swedish language changes over time. The system is designed as seen in Figure 4.1.

After retrieving a text, it is processed and most of the structure of the web text is removed. The text is separated into sections such as title and body. It is also annotated with the date when the text was published, this is the essential attribute which we will be able to use to track words over time. Some parts of the web text structure is preserved, such as tags for bold font and emphasis. The reason for this is that these tags may carry a semantical meaning (e.g. “That is an **e-mail**”). As a final step the text is automatically annotated with part-of-speech tags and is tokenized using the Swedish word class tagger Tagger (Carlberger and Kann, 1999).

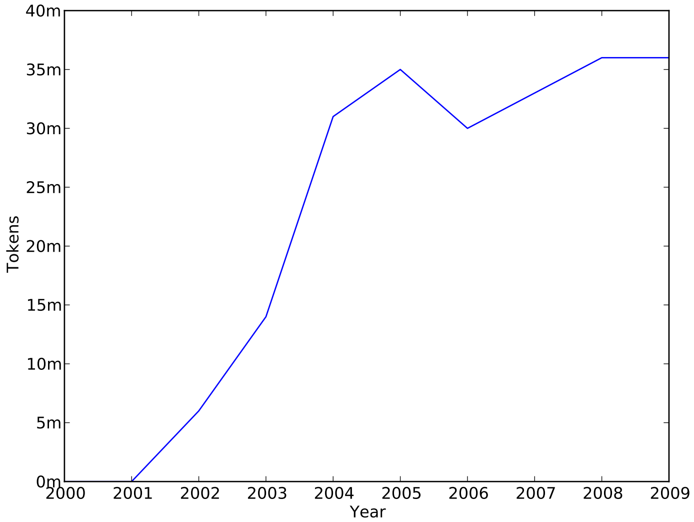


Figure 4.2. Number of tokens in the corpus per year from the year 2000 to 2009

While we are explicitly aiming for a large size Swedish corpus, using both blogs and newspapers from the Internet imposes restrictions on how wide of a time-span our corpus can cover. If we observe Figure 4.2 we see that we have a significant amount of data for the latter half of the last decade. But for the first half of the decade we only have a small amount of data. This is largely due to historical reasons, blogging first caught on from about 2004 and onwards and is still increasing. The publishing of newspaper articles on the web increased dramatically from 2001 to 2004, but appears to have stabilized. Prior to the year 2000 it is difficult to find temporally annotated resources and we are forced to rely on a time-span which may be wide but considerably narrower than we would like it to be. At the time of writing the corpus contains $225,752,550$ lexemes from $710,840$ documents, $1,833,626$ lexemes which are unique, it is most likely the largest temporally annotated Swedish corpus in existence.

Chapter 5

An Active System for Extracting Neologisms

Now that we have a large collection of data available we can focus on constructing a system for two purposes. Our system must be able to assist a lexicographer when deciding on whether a word is a neoligism or not and it should be able to extract and propose potential neologisms from the corpus.

A majority of the previous systems constructed for this task has been passive systems. What we mean by this notion is that they lay dorment and await user input, most commonly in the form of a text, produce output and return to a dorment state. Neologisms are in their very nature an active phenomena, what was a neologism several years ago may not be so today. The rules that applied for determining a neologism twenty years ago may be unsuitable for our language today. How can a system adapt to this reality?

5.1 Feedback and Result Cycle

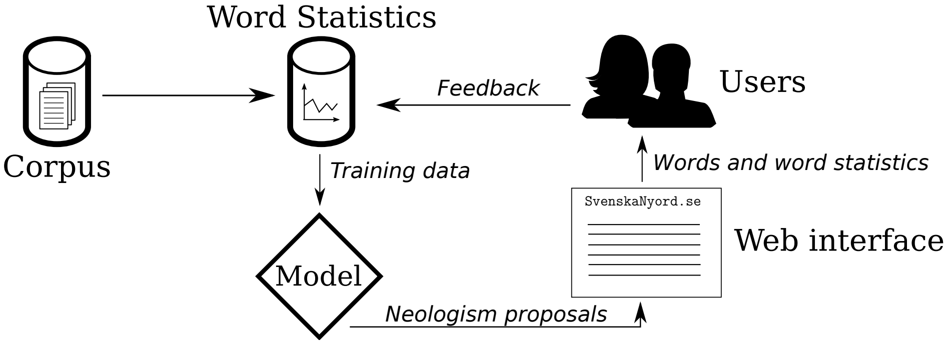
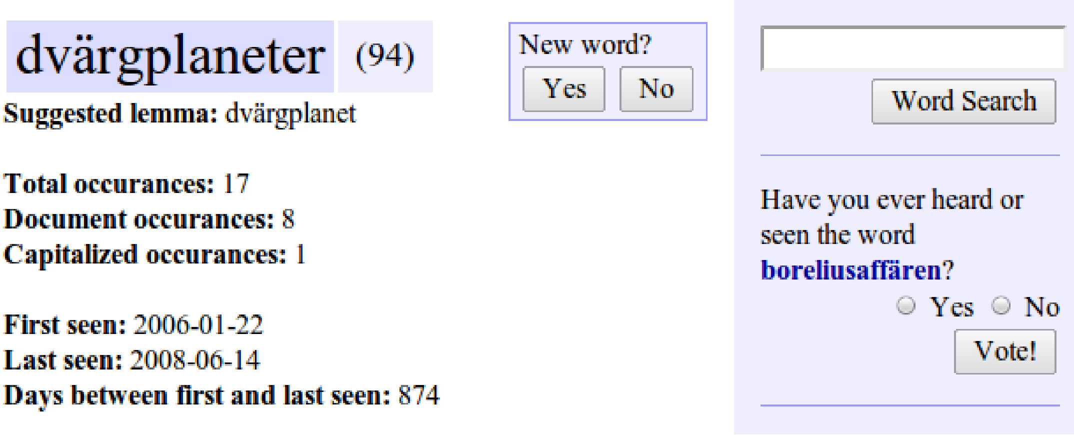


Figure 5.1. Active system design to extract neologisms which aims to gradually improve by incorporating user input

As previously argued by Janssen (2005a), any filtering applied poses a risk to exclude a neologism. Given the fact that an exclusion error is not recoverable and likely to make users unable to observe a certain neologism we may be forced to tolerate a high degree of noise.

A key difference in our system is that it does not employ any filters. All words from the corpus are considered, but instead of filtering the words they are ranked and given a score. Words with a low score are thought to be less likely to be neologisms and words with high score more likely to be neologisms. Using this simple tactic we avoid the problems caused by filtering, but we still have to consider the possiblity that true neologisms are given a low score and may go unnoticed.



dvärgplaneter (94)

Suggested lemma: dvärgplanet

Total occurances: 17
Document occurances: 8
Capitalized occurances: 1

First seen: 2006-01-22
Last seen: 2008-06-14
Days between first and last seen: 874

Have you ever heard or seen the word **boreliusaffären**?

Figure 5.2. The upper part of the web page of the word “dvärgplaneter” (dwarf planets)

To address this issue we leave the world of rule-based systems entirely. Instead of incorporating a single lexicographer to construct a collection of criteria for what is and what is not a neologism, we allow multiple users to partake in the effort. The system is visualized in Figure 5.1. Using a web-interface as seen in Figure 5.2 we allow any user to provide feedback on a word being a neologism. By basing our score on a machine learning model, we can use the feedback from the users in order to construct a new classifier and produce a new set of scores taking the new feedback into consideration.

In theory our classifier will improve over time in a manner similiar to Engelson and Dagan (1996) and Tsuruoka et al. (2008) which has previously applied this approach when annotation large quantities of data. Since the corpus contains well over a million words and is growing, assigning any human being to annotate the words is impossible. What our approach provides is a way to allow the system to learn more quickly by providing feedback for the cases which are most critical to improve our accuracy. If a non-neologism word is given a high score it will quickly be spotted by a user who will be able to point out to the system that the word does not deserve a high score. The system will the automatically re-consider all scores and assign a lower score.

A side note is that this system does in fact take multiple opinions on neologisms into consideration. Unlike previous systems which used a single lexicographer to judge the “newness” of a word. Our system can easily take opinions from varying sources and better reflect a large set of opinions on what constitutes a neologism.

5.2 Word Features

When using a machine learning method it is necessary to construct a model by extracting different features from the subject you wish to classify. In our case the features are mostly based on statistics from our corpus but also on external linguistic resources such as dictionaries and NLP tools. Something that is worth noting is that our system considers all lexemes in lower-case, but statistics is regarding upper-casing is collected.

- **Occurances:** Occurances in the corpus, number of documents the word is present in and number of capitalized occurrences
- **Occurances over time:** Occurances in the corpus and number of documents the word is present in, all with a three year look-back

Occurances is an indicator on how well established a word is. Functional words such as “and” will be more frequent than newer words or uncommon words. More important are the frequencies over time. It would be ideal to feed the system a fine-grained plot of the word occurrences over time, but our machine learning model only supports discrete steps, thus we are forced to decide upon an arbitary number of steps looking back from a given point in time. All of these features are normalized in relation to the total number of words in the corpus for during that time-span in order to compensate for the changes in the amount of data in the corpus as observed in Figure 4.2. The capitalized occurrences are used in relation to the true amount of occurrences, this is to enable the system to recognize proper names even though our system internally treats all words as being lower-case.

- **Points in time:** First time the word was observed in corpus and last time the word was observed in corpus
- **Age:** Number of days between the first time the word was seen in the corpus and the last time the word was seen in the corpus

The theory behind the points in time is that it will enable us to differentiate between old words and newer words. It would also enable us to see if a word has not been used for a very long period of time, thus avoiding words that are old and alien to most language users. The age can be used to determine if a word has only been used on a single occation or for a very short period of time. Since we are relying on news, there is a risk of a term being used in relation to a specific news story that only surfaces for a brief period of time. If the language users will not consider these short-lived neologisms to be valid neologisms the system will need a feature to rely on for this judgement.

- **Lexical cues:** Times the word is preceeded by “So-called”, marked with single or double quotes, marked with the HTML tags em, strong or italic

Our features for lexical cues builds on the research by Paryzek (2008) which we discussed in relation to previous research. We use some additional cues in relation to HTML tags since our corpus unlike Paryzeks contains these HTML tag information. It should be noted that “marked” means that a single word is assigned a tag or quoted and the neighbouring words are not (e.g. “It” is marked in the following sentence “There "it" is!”, bot not in “"There it is!"”). This restriction is done in order to avoid words contained in quoted phrases and large emphasized portions of texts as being more likely of being neologisms.

being interpreted as being an indication of a neologism.

- **Lexical resources:** The word is present in the Swedish and/or English Mozilla dictionary[1]
- **Spell checking:** The word passes the Stava (Domeij et al., 1994) spell-checker and the word does not pass the spell-checker and no word suggestions are provided

Finding free lexical resources for Swedish is not an easy task but the Mozilla Foundation[2] provides one for the Swedish language. Presence in a lexicon can be one indicator of a word having come of age.

More important are the spell checker attributes. Previous systems have had issues with misspellings and in order to remedy this we have chosen to utilize the best available spell checker for the Swedish language. We do not just use it to check if a word raises an error, we also exploit the fact that a spell checker can be stumped and return no suggestions for a correct word. This is usually done either for a completely new word or something which is not a word at all. In combination with other features we hope that this can identify potential neologisms.

5.3 User Interface

In order to assist users when extracting neologisms our system provides search functionallity and a word page for each word. The top part of a word page can be seen in Figure 5.2. The word page contains statistics and information similar to the word features mentioned in the previous secton. This gives the user a larger amount of metrics than the currently used tools such as searches in newspaper corpora and search-engines.

[1] `https://addons.mozilla.org/en-US/firefox/browse/type:3`

[2] `http://www.mozilla.org`

None of these older methods enables a user to see how metrics change over time. Most essential to our interface is the graph of how the word occurrences change in relation to its own occurences. An example of the neologism “Roasting” can be seen in Figure 5.3.

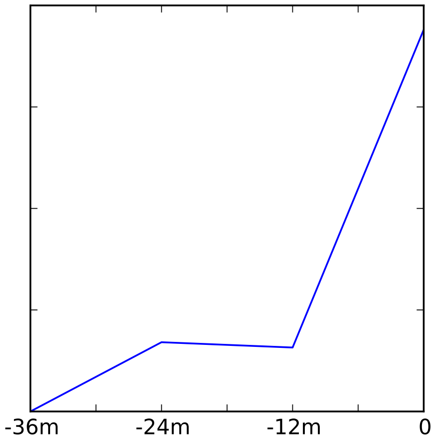
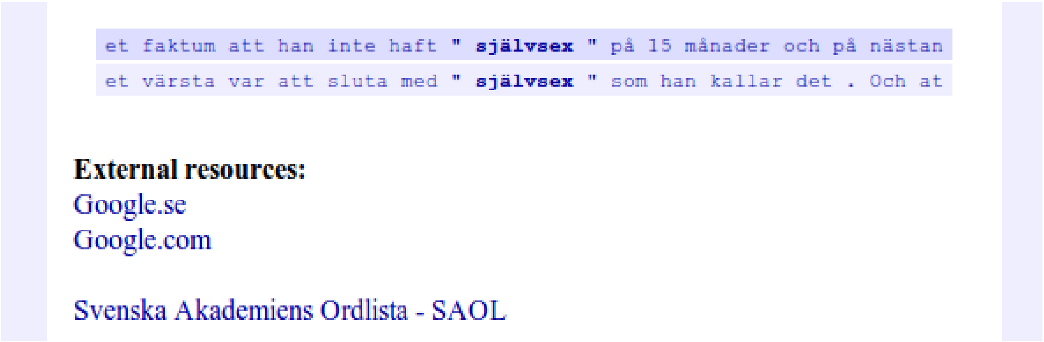


Figure 5.3. A plot of the frequency of the word “Roasting” from 36 months ago up to the current day, the plotted value is relative to the total number of lexemes

Neologisms have sharp rises in usage, while already established words remain fairly stable. We also compensate for the changes in the data amount in the corpus, otherwise even more common words would show a sharp increase if the size of the corpus was expanded.



et faktum att han inte haft " **självsex** " på 15 månader och på nästan

et värsta var att sluta med " **självsex** " som han kallar det . Och at

External resources:

Google.se

Google.com

Svenska Akademiens Ordlista - SAOL

Figure 5.4. Concordances supplied to the user to give the context in which the word is used. Below the concordances are external resources where the user can find additional information on a word

As seen in Figure 5.4 we supply a context for each word in order to allow insight into word-usage and meaning. In many cases seeing a single word may not be enough, especially if the user is not previously familiar with the word. Finally, at the bottom of the word page there are several links to other lexical resources which can aid the user in his or her decision. Such as looking up the word in large commercial dictionaries, on Wikipedia or using a search-engine.

Part III

Experiments and Results

Chapter 6

Training Data

Since our proposed system is based on machine learning methods and we use a Support Vector Machine (SVM) as an integral part of the system, it is necessary to provide positive and negative examples of neologisms as training data. From this training data it is then possible to produce a model to classify and rank all the words contained in our corpus.

It is the commonly agreed dogma in machine learning that the larger the amount of training data, the better the results. But for the problem of classifying neologisms we have to accept that no large amount of training data is readily available. We could select a random subset of the words contained in our corpus to be annotated by a group of volonteers, but this would most likely result in largely negative examples since non-neologisms are likely to be far more common than neologisms for our corpus.

Assigning an individual to select a subset of the words from the corpus for further evaluation would introduce bias on part of that individual and would not contribute towards solving our problem. In this chapter we will try to overcome these issues and acquire enough training data to train an initial model so that it can then be gradually improved by providing additional human feedback over time.

6.1 Acquiring Initial Training Data

Adressing the issue regarding positive training data we are reminded that the Swedish Language Council has previously released several annual lists of neologisms. Since our corpus is temporally annotated it is possible to generate statistics on the state of a given word in the corpus at any point in time. Seeing that the lists were released from the year 2000 and onwards, we can generate the statistics for the 31st of December for each year, which roughly corresponds to the time when the annual list was compiled and published, then use the word labeled as a neologism at that point in time as a positive example of an actual neologism. This approach is illustrated in Figure 6.1. While we do not expect the results of this initial training data to be good, we should be able to safely assume that it will perform better than chance at finding a neologism in the corpus data for further annotation.

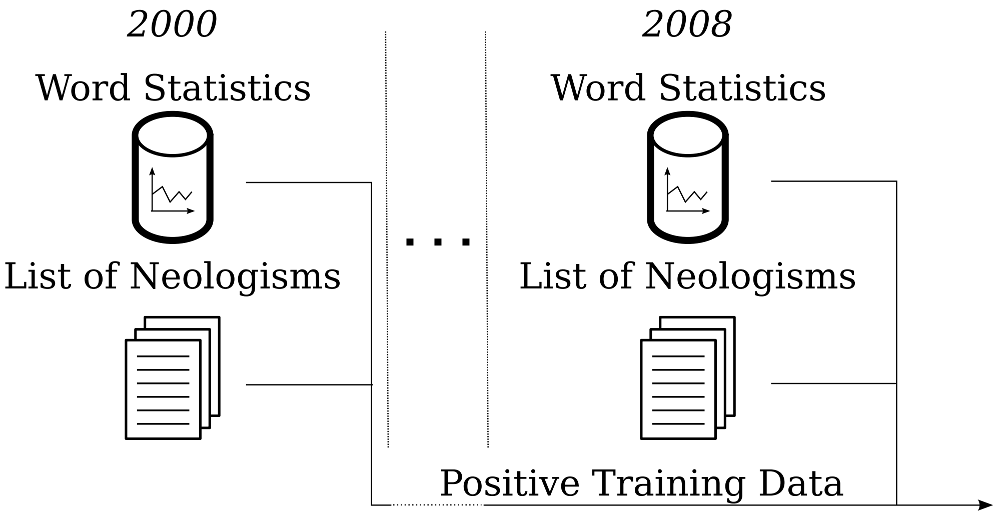


Figure 6.1. Procedure to produce initial training data based on previously released annual lists of neologisms and a corpus with temporal annotations

An issue concerning this approach is that only 45% (215 out of 478) of the neologisms from the annual lists of neologisms are present in the corpus. Since the majority of the corpus consists of texts collected from newspapers we are reminded of what was mentioned by Lindgren (2007), that newspaper texts may not be diverse enough to capture the language usage as a whole. Possibly the corpus may also be too small in size and in coverage, an issue which is discussed in Section 8.2 as a possible improvement. Another possibility is that the lists of neologisms may contain a considerable amount of words which are seldom used or may only be used occassionally. Further research into this matter is necessary to pass a final judgement but this is beyond the scope of this thesis.

As negative training data a random sample of $2,500$ words from the Swedish PAROLE[1] corpus was used. The reason for choosing PAROLE is that due to its high age it is unlike that it will contain neologisms. The material contained in the corpus is at youngest from 1997 but ranges back as far as 1981. Even in the case that the random sample contains neologisms, it is in accordance with machine learning theory that end result should not be significantly impaired as long as the majority of examples are accurate.

The training data was fed into an SVM and a model was trained and optimized in accordance with the general method provided by Hsu et al. (2003). The feedback from this initial model will be analyzed and discussed for the remainder of this chapter.

[1] `http://spraakbanken.gu.se/parole/` *(in Swedish)*

6.2 Acquiring Diverse Training Data

After training our initial model we now have a way to rank the words contained in our corpus. Since we have previously argued that neologisms are best reflected in the opinion of a diverse set of language users, we need to find a way to gather data which is more diverse than the neologism lists published by the language council.

For this task twelve anonymous annotators were recruited. They were chosen from different age-groups, occupations, educational backgrounds and geographic locations. We assume based on our theoretical background that the views of these language users will better reflect what an average language user would consider to be a neologism.

The model trained using the initial training data, classifies $142,961$ out of $1,833,626$ words as being neologisms. That all of the classified words are neologisms is of course unreasonable, but the process has hopefully excluded a considerable amount of non-neologisms from our candidates. Under this assumption a random sample of $2,000$ of the words classified as neologisms were selected.

Batches of 100 words were then supplied to the anonymous annotators and annotated using the interface described in Appendix B. The annotators were instructed to rely on their own proficiency in the Swedish language rather than relying on external sources to base their annotation on. To the knowledge of the author, there exists no prior research on how neologisms are perceived by language users, this annotation effort was then a chance to further investigate how well our previously discussed theories and assumptions co-relate with the annotated data. Thus the set of annotations were expanded beyond classifying whether or not a word was a neologism.

- **Neologism:** The annotator considers the proposed neologism to be an existing word and to be a new addition to the language
- **Word:** The annotator considers the proposed neologism to be an existing word, but a word with no sense of “newness”
- **Not a word:** The annotator considers the proposed neologism not to be an existing word
- **Old word:** The annotator consiers the proposed neologism to be an existing word, but the word feels old and it would feel unnatural to use it

The motivation for including the two additional word types “Not a word” and “Old word” is to better understand where our classifier is making mistakes. One fear is that like previous systems we present a large amount of non-words, this would encourage us to revise our model and perhaps include more features reflecting word structure. If many truly old words are included it may be because our corpus includes words which has not been used for many years, these words are even likely not to be present in the PAROLE corpus and could then be lacking as negative examples in the tranining data. Pointing out a flaw in our decision to use a random sample from PAROLE as negative training data.

In addition three binary attributes somewhat unrelated to the word types were included.

- **Knowledge:** The annotator has prior knowledge of the proposed neologism, that is, the annotator has seen, heard or used this word prior to annotating it
- **Misspelling:** The annotator considers the proposed neologism to be a misspelling of an existing word
- **Skip:** The annotator is not confident in judging the proposed neologism and wishes to ignore it

"Knowledge" will give us a more clear understanding of how alien the presented neologism is to the language user. The "Misspelling" attribute is included since previous systems have had issues with misspellings being classified as neologisms. As seen in Section 5.2 we have taken precautions towards these kinds of errors by incorporating features that should better enable our SVM to differentiate neologisms from misspellings. If these approaches needs to be refined, this annotation will give us a clear indicator. "Skip" is a simple convenience attribute for the annotators so that they can avoid making annotations that they are not comfortable with.

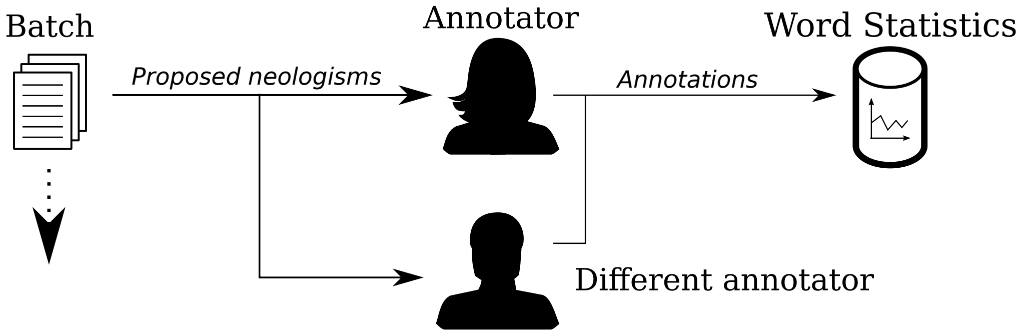


Figure 6.2. The annotation procedure

As depicted in Figure 6.2, each batch of words is annotated by two independent annotators, this way we will cover a larger portion of the words than if we assigned more than two annotators per batch. Not using a single annotator per batch will enable us to analyze the data in regard to annotator agreement, which is an important aspect in relation to our theoretical background regarding neologisms. A downside is that when performing an annotation it would be greatly beneficial if a large number of language users could give their opinion on a neologism. If the annotator was willing to do so, he or she was allowed to annotate an additional batch after annotating his or her first batch. This enabled us to cover an even larger amount of words and still limit the amount of bias introduced by allowing a single annotator to annotate a large amount of the words.

In the next section we will discuss a large portion of the results from the annotation. Some of the discussion is located in Appendix C, since it may be of interest for future annotation efforts but is not relevant to the purpose of this thesis.

6.3 Analysis of Annotated Data

We should first focus on the assigned word types and discuss the implications. The word type annotations are summarized in Table 6.1 and Figure 6.3.

We can see that the vast majority of the output consists of Swedish words, while not surprising this is of course not ideal for our classifier. Possibly this is due to a lack in diversity on part of the negative training data from PAROLE. We could explain the large amount of non-words using the same argument. None of the negative examples of neologisms extracted from PAROLE were non-words and we should perhaps have chosen to use PAROLE in combination with a source of non-words in order to acquire a better initial classifier.

However, we should now have a considerable amount of training data for these cases and our next classifier should stand a better chance at avoiding the same mistakes. This can be seen as a positive side-effect of our annotation effort.

Word type	Words assigned
Word	1292
Neologism	213
Not a word	218
Skip	248
Old word	29
Total:	2,000

Table 6.1. The amount of words assigned with a specific type in the annotated data

A more urgent observation is what we observe in Table 6.2 and Figure 6.4. The level of agreement between the annotators is remarkably low and this has implications on the performance which we can expect from any automated system for extracting neologisms. Any system which is subject to evaluation must then somehow address the issue that annotators will disagree in regard to its results.

If we remind ourselves of the discussion regarding languages and neologisms in relation to linguistic theory in Part I, we expect the notions of language users in regard to what is and what is not a neologism to differ. But our annotated data shows that the average user will only agree on what is a neologism in one out of five cases. In the previous section we discussed the reason for only having two annotators

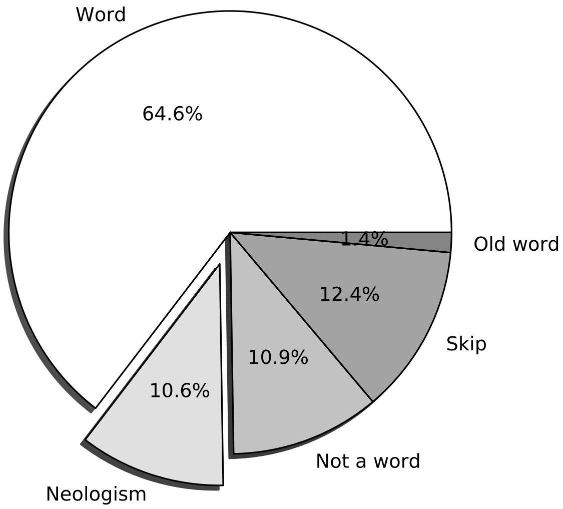


Figure 6.3. Distribution of the word types from Table 6.1

to annotate each batch, but these results suggests that further investigations into this matter with more annotations per word may shed light on how neologisms are perceived by the language users.

Type	Amount
Neologisms	178
Agreed upon neologisms	35

Table 6.2. The amount of words assigned as neologisms by either annotator and the amount assigned by both annotators as being neologisms

We now combine the previous training data extracted from the lists of neologisms and our annotated data to train our final classifier. The results from this classifier will be discussed in the Chapter 7.

6.4 Summary

In this chapter we have addressed the lack of available neologisms to use as training data for our system. We first exploited the fact that several annual lists of neolo-

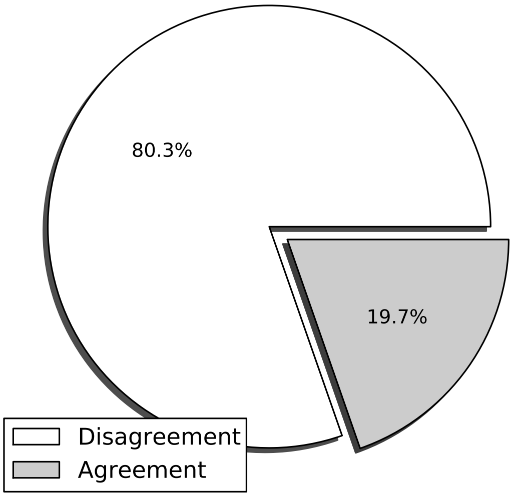


Figure 6.4. Agreement among annotators regarding neologisms from Table 6.2

gisms has been published by the Swedish Language Council. We did this using the lists in combination with our temporally annotated corpus and could thus acquire the features of a word from the time when the list was published and compiled.

Using this positive traning data in combination with a random sample from the PAROLE corpus as negative examples we trained a model that we then used to classify all the words in our corpus as either neologisms or non-neologisms. From the words classified as neologisms we then took a random sample to be annotated by twelve volonteers.

The annotations showed that our initial training data was insufficient and lacked the coverage needed to exclude many non-neologisms. This was especially prevalent in regard to ordinary words but also in relation to non-words being present among the proposed neologisms. We also noted that the level of agreement among annotators on what is and what is not a neologism was remarkably low. Only one out of five neologisms were agreed upon. We discussed the implications of these results and then proceeded with training a final classifier by combining our initial training data with the annotated data.

Chapter 7

Results

The result of our efforts is a system which can both aid users and to a certain extent automate the process of neologism extraction. We have produced a highly modular system in which components can be replaced with ease. Since we are employing machine learning methods our model can also be changed to experiment with different viewpoints and to incorporate future research from other sources.

7.1 Aiding Lexicographic Extraction of Neologisms

When aiding users to extract neologisms our system supplies a wide arrange of word features as a by-product of our need for word features for our machine learning methods. These features enables a user to better judge a word in regard to it being a neologism than current methods such as web-searching.

Temporal changes in word usage are captured and displayed in an intuitive manner and the user is also provided with several tools to assist him or her in her work.

7.2 Automatic Extraction of Neologisms

Evaluating the performance of our system is as we have argued a difficult task. As we expect our system to improve over time and that our training data imposes further restrictions on what conclusions we may draw at this stage.

Using our training data from the annotated data we once again train a classifier and let five anonymous annotators assign word types to five batches. The results of these annotations can be observed in Table 7.1 and Figure 7.1.

What we can see is that while the ratio of neologisms has remained static, the number of non-words has increased to an even higher level. These results may at first seem discouraging but prompts a revision of our current model. While 10% may seem a very low level of accuracy, it is on-par or better than the accuracy of all previous systems and methods. The task of neologism extraction has proven to be hard and the underlying theory supports this. Rather than as an initial failure this

Word type	Words assigned
Word	254
Neologism	53
Not a word	74
Skip	87
Old word	32
Total:	500

Table 7.1. Output of the final classifier annotated by word types

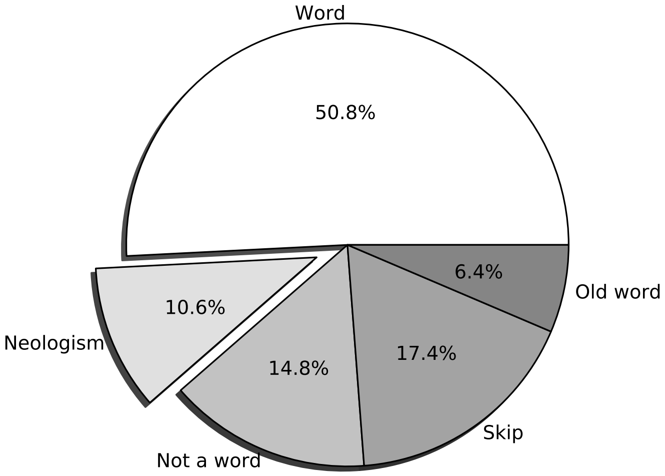


Figure 7.1. Distribution of the word types from Table 7.1

system proves that we can incorparate a wide array of features in order to better model a phenomena which is both complex and subjective.

7.3 Results Compared to the Annual List of Neologisms of 2009

As an interesting comparison of the results of our system we relate it to the annual list of neologisms of 2009 from the Swedish Language Council. The list of 31 words is presented in Table 7.2. For each word in the list we also have the score that was assigned by our final classifier. The score assigned by the system ranges from 0 to 100, a high score is to be seen as an indication of the word being a neologism. If the word is not present in the corpus it is marked with a dash.

Word	Assigned score
Alfanummer	18
Bilsurfa	23
Chefsnappning	43
Chippa	0
Fiskpedikyr	44
Frimester	7
Fröbomba	7
Fuldelning	21
Följare	35
Grindstad	22
Gågging	11
Hemester	53
Kvittra	7
Könskonträr	15
Laddstolpe	65
Mobilroman	7
Norsk karaoke	-
Oresebyrå	-
Prokotta	15
Sitskate	20
Slidkrans	40
Spikmatta	3
Sporta	0
Sprita	3
Stjärnfamilj	-
Stuprörspolitik	19
Svemester	19
Svininfluensa	21
Tvittra	15
Twittra	25
Yrkessåpa	-

Table 7.2. The annual list of neologisms of 2009 from the Swedish Language Council, each word assigned with a score from our final classifier

We can see that none of the neologisms from the Swedish Language Council is assigned a considerably high score. Compare this list to the words in Table 7.3.

These are the words from the corpus that was assigned the highest score by our final classifier.

Word	Assigned score
Allmänfattlig	100
Antager	100
Bedrager	100
Bekehrung	99
Besvuren	100
Burgteatern	99
Cicisbeo	99
Elsassisk	100
Epidermis	100
Förtaga	100
Gruvsam	100
Grälaktig	100
Handhava	100
Hinderlig	99
Indragit	99
Ingenium	99
Landeskriegsfürsorgeamt	99
Lupit	100
Observationsmaterial	100
Phädra	99
Pjollrig	99
Polara	100
Rättsinnig	100
Skyldskap	100
Spektralanalytisk	100
Subordination	100
Tackoffer	100
Utdraga	100
Utgallra	99
Uttränga	100
Verop	100

Table 7.3. The 31 words in the corpus that was assigned the highest score

Most of the words in this list can never be considered neologisms despite of their high score. When observing these word attributes we find that a majority of them have been marked with lexical cues. It is very possible that due to our very limited training data we have received a classifier which is biased in regard to lexical cues.

But we must remember that we set out to construct a system that would most likely not replace human efforts. If we go further down the list of words in the corpus sorted by score, we find “organturismen” (the organ tourism) with a score of 98, “whistle-blower” (whistle-blower) with a score of 98 and “yttrandefrihetsfundamentalism” (free-speech fundamentalism) with a score of 97. These words may in fact the considered to be neologisms, but it is still necessary for a human being to partcially participate in the process of finding them.

Chapter 8

Conclusions

In this thesis we have investigated and documented the efforts to construct an automatic system for extracting neologisms. We have also described previous approaches, the underlying linguistic theory and described how our system relates to both. We also presented linguistic theory and observation to support our standpoint that neologisms are a largely subjective notion. Our final system can potentially be trained to capture either a very general notion of what constitutes a neologism or a narrow notion based on the training data that is provided. This feature is unique for our system in relation to previous systems for neologism extraction.

8.1 Fulfillment of Purpose

We consider this thesis to have fulfilled its purpose. We have constructed and presented a system capable of extracting neologisms from a large temporally annotated corpora. While our final results are not on-par with the results of a human extractor, we find that our system captures some aspects of neologisms. Our system is also designed to improve over time and by reconsidering the model and providing additional feedback we can expect the results to improve over the coming years.

Our system is also capable of assisting a user in the extraction of neologisms. Acting as a tool that provides information which was previously unavailable to the user such as changes in word attributes over time. This is a great boon for any user seeking to base his or her judgement on facts based on actual word usage.

Our final conclusion is that neologism extraction is a difficult task and that simplistic non-statistical approaches are unlikely to produce a system capable of grasping a language-phonomena of this complexity.

8.2 Future Work

As an extension we would like to expand the size of the corpus. The reason for using newspaper texts in linguistic research is probably mostly due to historical reasons.

Newspaper texts are readily available but they are not a good source of texts from average language users. For this, a more diverse range of blogs should be mined for data and we should also consider less formal sources such as micro-blog feeds and e-mail conversations.

An interesting possibility to investigate is whether word co-locations can be used to capture semantical changes to a word. Although it is daring to suggest that a lexical feature such as co-locations can capture something as complex as word semantics it is an interesting thought.

A problem with our approach using an SVM is that it only allows for binary classification. Supported by our theory we can argue that “newness” is not binary in nature but rather continuous. Training data should then not be labeled as positive or negative, but the annotator should rather be asked “How new do you feel that this word is?” and be supplied with a continuous scale of “newness”. Employing other machine learning methods and performing additional annotation of words from our corpus should enable us to train the classifier in a new manner and perhaps produce better results and be more in line with our theoretical view on neologisms.

Acknowledgements

No work is solely based on the efforts of one individual, so the author would like to acknowledge the input and support provided over the last year. As in all acknowledgements the author fears that someone may have been forgotten in the process of writing, if this is the case, rest assured that this is a reflection of the short-comings of the author rather than of the quality of your contributions.

First of all, I would like to express my gratitude to the members of the Tsujii Laboratory at the University of Tokyo, for their support and love during my stay at the laboratory from April 2009 to March 2010. In particular to Professor Jun'ichi Tsujii for having confidence in me and my efforts, thus enabling me to grow as an individual and as a researcher.

Since my personal background foremost has been in computer science, I lacked insight into the field of linguistics and found the ideas and discussions with Luke McCrohon, also of Tsujii Laboratory, to be most rewarding. I hope that we may get that paper we have been working on accepted some time soon and continue our co-operation in evolutionary linguistics which he has introduced me to.

I owe the twelve anonymous annotators a lot of gratitude for putting up with my demands and providing me with data justifying my approach. The hours spent by you are forever a part of this thesis.

Although the thesis was carried out at both the Royal Institute of Technology in Stockholm and at the University of Tokyo, a great deal of support was provided by the Swedish Language Council. Financial support was provided by Ebba Danelius Stiftelse, which provided me and my supervisor with financial means to support ourselves and acquire the technical equipment necessary to house the software used for this thesis.

Last but not least, I would like to sincerirely thank Professor Viggo Kann for his endless patience and wisdom. I can not imagine that without him as a supervisor I could ever have felt the confidence to stride onwards and look upon this task with new eyes. What was once a cautious and naive approach to an interesting problem soon was transformed into a grand scheme, hoping to leverage well established methods in a new and unique fashion. Thank you for being the best supervisor I could ever have hoped for.

Bibliography

G. Andersen. Assessing algorithms for automatic extraction of Anglicisms in Norwegian texts. *Corpus Linguistics*, 2005.

K. Bjarnadóttir and E. Rögnvaldsson. Automatiske metoder til excerpering af nye ord, 2007. Seminar om Sprogrøgt, Sprogpolitik og Sprogteknologi i Norden.

M.T. Cabré. *Terminology: theory, methods, and applications*. John Benjamins Publishing Company, 1999.

J. Carlberger and V. Kann. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 1999.

B. Cartoni. Dealing with unknown words by simple decomposition: feasibility studies with Italian prefixes. *Actesde LREC*, 2006.

W. Croft. *Explaining language change: an evolutionary approach*. Pearson Education, 2000.

R. Dawkins. *The extended phenotype*. Oxford University Press Oxford, 1982.

D.C. Dennett. *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster, 1995.

R. Domeij, J. Hollman, and V. Kann. Detection of spelling errors in Swedish not using a word list en clair. *Journal of Quantitative Linguistics*, 1(3):195–201, 1994.

U.C. Eiken, A.T. Liseth, H.F. Witschel, M. Richter, and C. Biemann. Ord i Dag: Mining Norwegian Daily Newswire. *Lecture Notes in Computer Science*, 4139: 512, 2006.

S.P. Engelson and I. Dagan. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 319–326. Association for Computational Linguistics Morristown, NJ, USA, 1996.

J. Halskov. Itstøttet excerpering af sproglige nydannelser, 2007. Seminar om Sprogrøgt, Sprogpolitik og Sprogteknologi i Norden.

M. Hassel. Internet as Corpus: Automatic Construction of a Swedish News Corpus. In *Proceedings of Nodalida*, 2001.

K. Hofland. A self-expanding corpus based on newspapers on the web. Technical report, HIT Centre, 1998.

K. Hofland. Ekserpering av nyord fra norske webaviser, 2007. Seminar om Sprogrøgt, Sprogpolitik og Sprogteknologi i Norden.

C.W. Hsu, C.C. Chang, C.J. Lin, et al. A practical guide to support vector classification, 2003.

M. Janssen. NeoTrack: semi-automatic neologism detection. Papel presentado en la conferencia XXI APL, Lisbon, 2005a.

M. Janssen. Orthographic Neologisms: selection criteria and semiautomatic detection. 2005b.

B. Lindgren. Nyordsexcerpering, 2007. Seminar om Sprogrøgt, Sprogpolitik og Sprogteknologi i Norden.

P. McFedries. The internet ecology. *IEEE Spectrum*, 40(4):68–68, 2003.

P. Paryzek. Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisicae*, 16:163–181, 2008.

A. Rey. *Essays on terminology.* J. Benjamins, 1995.

Svenska Akademien. Svenska Akademiens Ordlista. 2006.

Svenska Språkrådet. *Nyordsboken.* Norstedts Ordbok, first edition, 2000.

Svenska Språkrådet. *Nyord i Svenskan från 40-tal till 80-tal.* Norstedts Ordbok, third edition, 2001.

Y. Tsuruoka, J. Tsujii, and S. Ananiadou. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC bioinformatics*, 9, 2008.

Appendices

Appendix A

Svenska Nyord Corpus XML-format

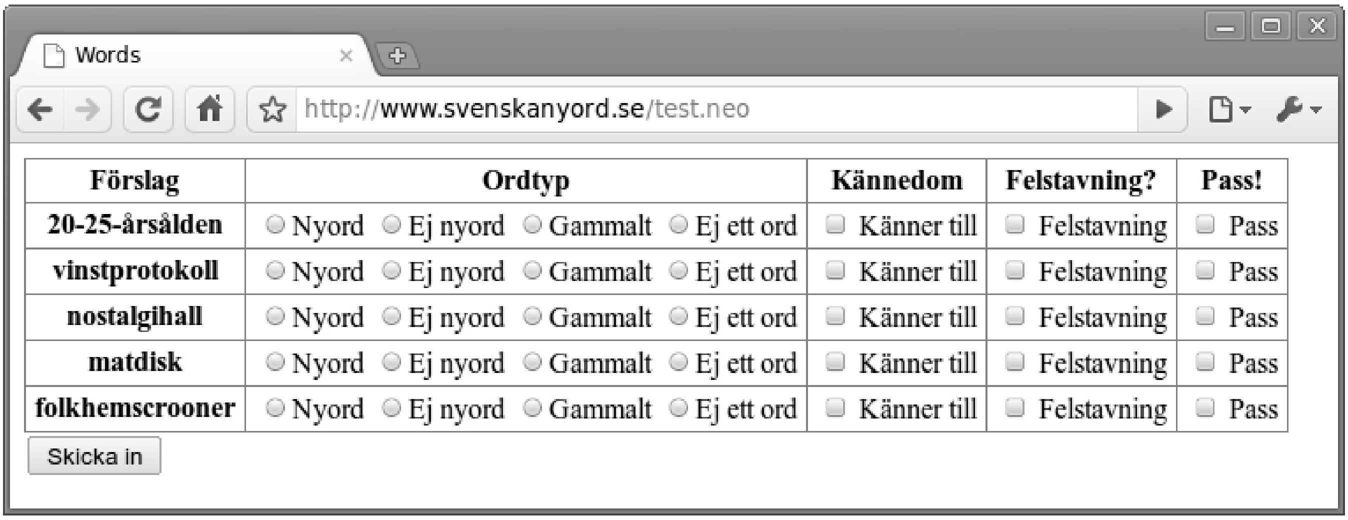
This appendix aims to briefly explain the Svenska Nyord corpus format. The format attempts to reflect the structure of an article or blog post. By using an XML-format it can use the well-established rules for escaping characters and can be accessed using efficient tools and libraries for XML processing. For a brief example of an article from the corpus, please see Figure A.1.

```
<?xml version="1.0" ?>
<article author="Carolina Neurath" checksum="746503532" extracted="↩
    2010-01-13 12:42 UTC" inserted="2010-01-13 16:41 UTC" published="↩
    2010-01-12 16:30 UTC" url="http://www.svd.se/naringsliv/nyheter/genii-↩
    om-saabaffaren-vi-har-till-lunch-tid-pa-oss_4078101.svd">
    <title>
        <token html_tags="" is_word="False" lemma="&quot;" new_word="False"↩
             pos="pad" token_type="citation">
            &quot;
        </token>
        <token html_tags="" is_word="True" lemma="vi" new_word="False" pos=↩
            "pn.utr.plu.def.sub" token_type="simple word">
             &quot;Vi&quot;
        </token>
        <!-- Lines removed to fit the page -->
    </title>
    <content>
        <token html_tags="strong" is_word="True" lemma="budgivarna" ↩
            new_word="False" pos="nn.utr.plu.def.nom" token_type="simple ↩
            word">
            Budgivarna
        </token>
        <!-- Lines removed to fit the page -->
    </content>
</article>
```

Figure A.1. An example of an article from the corpus which has had most lines removed to fit the the page while maintaining the structure of a corpus entry

Appendix B

Training Data Annotation Application User-interface



Förslag	Ordtyp				Kännedom	Felstavning?	Pass!
20-25-årsålden	○ Nyord	○ Ej nyord	○ Gammalt	○ Ej ett ord	□ Känner till	□ Felstavning	□ Pass
vinstprotokoll	○ Nyord	○ Ej nyord	○ Gammalt	○ Ej ett ord	□ Känner till	□ Felstavning	□ Pass
nostalgihall	○ Nyord	○ Ej nyord	○ Gammalt	○ Ej ett ord	□ Känner till	□ Felstavning	□ Pass
matdisk	○ Nyord	○ Ej nyord	○ Gammalt	○ Ej ett ord	□ Känner till	□ Felstavning	□ Pass
folkhemscrooner	○ Nyord	○ Ej nyord	○ Gammalt	○ Ej ett ord	□ Känner till	□ Felstavning	□ Pass

Figure B.1. Web-interface used to annotate suggested neologisms provided by the model from section 6.1, to generate test data as described in section 6.2

The interface provides the following columns. Suggestion, word type, known, misspelling and skip. The suggestion column contains the suggested neologism to be annotated. Word type provides four alternatives: neologism, not a neologism, old and not a word. Known is checked if the annotator is familiar with the word. Misspelling if the annotator considers the word to be a misspelling. Skip is provided since in some cases the annotator may wish not to pass any judgement on the word. After all words has been annotated the annotator can submit the annotations to the system by pressing the send button.

Appendix C

Additional Annotation Data Analysis

The following appendix covers the data collected but not covered by the discussion regarding annotation data in Section 6.3. While being interesting when considering neologisms it is not directly relevant to the thesis, it is provided to the reader for future reference and research.

For this part of the thesis we should bring to the attention of the reader the feedback given by the annotators of their experience performing the annotation. As seen in Figure B.1 of the user-interface utilized for the annotation effort, no context is given for each word. Although annotators could pass judgement and understand the meaning of words previously not observed, they found it difficult to do so. As we see in Figure C.1, as many as 12% of the proposed neologisms were skipped. This is also observed in Figure C.2, it is unlikely that as much as 42% of the words were previously unknown to the user, rather than being difficult to recalling a word being given no context. To remedy this it was requested that a future system would provide some additional context in which the word was introduced. For future annotation efforts this is an issue that must be considered.

Unlike most of the data provided on the annotated data regarding misspellings in Table C.2 does not contain a total of 2,000 words. The reason for this is that some of the proposed words were classified as not being words or skipped. These words has been left out of the table to provide an accurate picture of amount of misspellings.

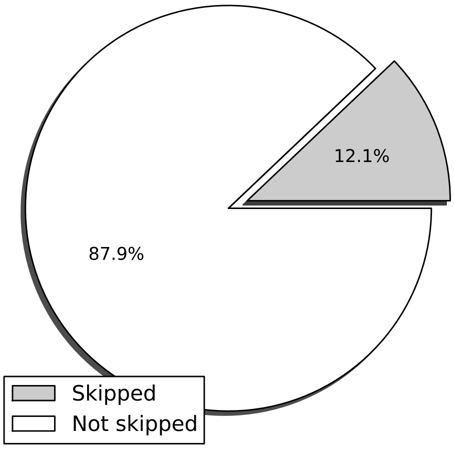


Figure C.1. Distribution for the 248 out of $2,000$ words, which were skipped by the annotators

Type	Amount
Prior knowledge	$1,155$
No prior knowledge	845
Total:	$2,000$

Table C.1. The amount of words marked by the annotators as being words which the annotator had prior or no prior knowledge of

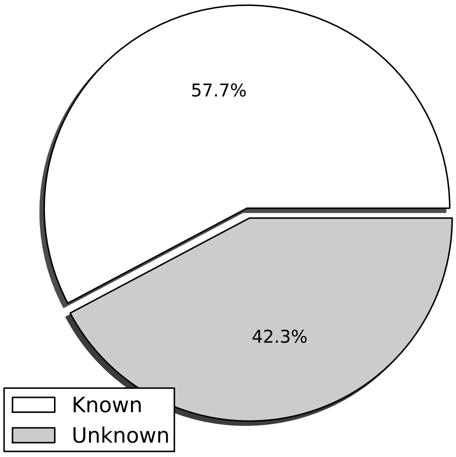


Figure C.2. Distribution of the data in Table C.1 regarding prior knowledge of the proposed neologism

Type	Amount
Misspelling	113
Not a misspelling	1,421
Total:	1,534

Table C.2. The amount of words marked as misspellt by the annotators

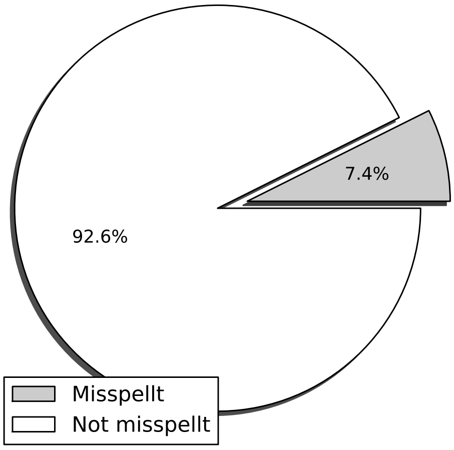


Figure C.3. Distribution of the data in Table C.2 regarding whether the proposed neologism was a misspelling of an existing word

Glossary

anglicism a word borrowed from English into another language or a word in another language that has changed its form or meaning due to the influences from the English language.

computer science is the scientific study of the theoretical foundations of information and computation.

de-facto is an expression to denote that something is commonly accepted as being true rather than established by authority.

ideogram is a graphic symbol that represents an idea or concept rather than a sound as is done in English (e.g. the Chinese character for east 東 or school 学).

lexeme is an abstract term used in linguistics, that roughly corresponds to the more informal notion of a *"word"*.

lexicographer is a person concerned with compiling, writing and editing dictionaries.

linguist is a person involved in the study of linguistics.

linguistics is the scientific study of natural languages.

metric is a measurement which can be applied to an entity, for example physical entities has a weight which is metric.

natural language or human language is a term used to differentiate naturally occuring languages such as English or Japanese to constructed and formal languages such as George Orwell's Newspeak and the language of mathematics.

Natural Language Processing is a field of computer science and linguistics which is concerned with the interaction between natural languages and computers.

neologism is a newly coined word or a word which has recently been given a new meaning. Informally a neologism may be refered to as a *“new word”*.

NLP Natural Language Processing.

semantics is the study of meaning and is considered a field within linguistics.

string is a computer science term that refers to a sequence of characters.

Support Vector Machine is a method in computer science used to construct and train a computer model capable of classifying data according to certain patters (e.g. to determine if a picture contains faces).

SVM Support Vector Machine.

training data is a term used for data annotated by human beings as being of a certain kind. This data is then assumed to be examples from which a machine learning algorithm (e.g. a Support Vector Machine) can train a model to classify new unknown data.

zeitgeist is a German expression for “the spirit of the times” similar to the English word “trend”.

TRITA-CSC-E 2010:017
ISRN-KTH/CSC/E--10/017--SE
ISSN-1653-5715

www.kth.se